Quality Gain Analysis of the Weighted Recombination Evolution Strategy on General Convex Quadratic Functions

Youhei Akimoto, Anne Auger, Nikolaus Hansen

Quality Gain Analysis of the Weighted Recombination Evolution Strategy on General Convex Quadratic Functions

Youhei Akimoto
Shinshu University
380-8553 Nagano, Japan
y_akimoto@shinshu-u.ac.jp

Anne Auger
Inria, Research Centre Saclay
91120 Palaiseau, France
anne.auger@inria.fr

Nikolaus Hansen
Inria, Research Centre Saclay
91120 Palaiseau, France
nikolaus.hansen@inria.fr

ABSTRACT

We investigate evolution strategies with weighted recombination on general convex quadratic functions. We derive the asymptotic quality gain in the limit of the dimension to infinity, and derive the optimal recombination weights and the optimal step-size. This work is an extension of previous works where the asymptotic quality gain of evolution strategies with weighted recombination was derived on the infinite dimensional sphere function. Moreover, for a finite dimensional search space, we derive rigorous bounds for the quality gain on a general quadratic function. They reveal the dependency of the quality gain both in the eigenvalue distribution of the Hessian matrix and on the recombination weights. Taking the search space dimension to infinity, it turns out that the optimal recombination weights are independent of the Hessian matrix, i.e., the recombination weights optimal for the sphere function are optimal for convex quadratic functions.

CCS Concepts

•Mathematics of computing → Bio-inspired optimization; •Theory of computation → Theory of randomized search heuristics;

Keywords

Evolution strategies; recombination weights; optimal step-size; quality gain analysis; general convex quadratic function

1. INTRODUCTION

Evolution Strategies (ES) are bio-inspired, randomized search algorithms to minimize a function $f : \mathbb{R}^N \to \mathbb{R}$ in continuous domain. The most commonly used variant of evolution strategies, namely the covariance matrix adaptation evolution strategy (CMA-ES) [17, 18], is one of the state-of-the-art randomized search algorithms for black-box continuous optimization. It maintains a multivariate normal distribution from which candidate solutions are sampled. The parameters of the multivariate normal distribution are updated using the candidate solutions and their function value ranking. Due to its population-based and comparison-based nature, the algorithm is robust and effective on non-convex and rugged functions.

Often the performance evaluation of evolutionary algorithms is based on empirical studies. One of the reason is that mathematically rigorous analysis of randomized algorithms are often too complicated due to the comparison-based and population-based nature and the complex adaptation mechanisms. To perform a rigorous analysis we often need to simplify some algorithmic components. However, theoretical studies can help our understanding of the behavior of the algorithms, provide optimal scenario that may not be empirically recognized, and reveal the dependency of the performance on the internal parameter settings. For example, the recombination weights in CMA-ES are selected based on the mathematical analysis of an evolution strategy [4][1]. Moreover, the optimal rate of convergence of the step-size is used to estimate the condition number of the product of the covariance matrix and the Hessian matrix of the objective function, which a recent variant of CMA-ES exploits for online selection of the restricted covariance matrix model [2].

Analysis based on progress rate or quality gain are among the first theoretical studies of evolution strategies that were carried out (see [11] for historical results of different variants of evolution strategies). Progress or quality gain measures the *expected progress in one step* (measured in terms of norm for the progress rate and objective function for the quality gain). Simplifying assumptions are then made to be able to derive explicit formula. Typically on spherical functions, it is assumed that the step-size times the dimension divided by the norm of the mean of the sampling distribution is constant. This allows to derive quantitative explicit estimates of the progress rate for the dimension N large that are correct in the limit of N to infinity (see (10)). Those analysis are particularly useful to know the dependency of the expected progress on the parameters of the algorithm such as the population size, number of parents, and recombination weights. Based on these results, one can derive some

[1]The weights of CMA-ES were set before the publication [4] because the theoretical result of optimal weights on the sphere was known before the publication.

FOGA '17, January 12-15, 2017, Copenhagen, Denmark
© 2017 ACM. ISBN 978-1-4503-4651-1/17/01…$15.00
DOI: http://dx.doi.org/10.1145/3040718.3040720

optimal parameter setting, in particular the recombination weights [4].

The progress rate on the sphere is linked to the convergence rate of "real" algorithms, that is implementing a proper step-size or/and covariance matrix adaptation. First, it is directly related to the convergence rate of an "artificial" algorithm where the step-size is set proportionally to the distance to the optimum (see [8] for instance)[2]. Second, the convergence rate of this artificial algorithm for a proper choice of proportionality constant gives a bound on the convergence rate of step-size adaptive algorithms. For $(1+\lambda)$ or $(1,\lambda)$ ESs the bound holds on any function with a unique global optimum, that is, a step-size adaptive $(1 \overset{+}{,} \lambda)$-ES optimizing any function f with a unique global optimum will not achieve a convergence rate faster than the convergence rate of the artificial algorithm with step-size proportional to the optimum on the sphere function [6, 20, 22][3]. For algorithms implementing recombination, this bound holds not on any f but on spherical functions [6, 20]. While analyzing the convergence and convergence rate of "real" algorithms is generally quite intricate, there is a simple connection between progress rate analysis and convergence analysis of step-size adaptive evolution strategies: on scaling-invariance functions (with optimum in zero without loss of generality (w.l.g.)), the mean vector divided by the step-size is a Markov chain whose stability analysis leads to the linear convergence of the algorithm [9]. In progress rate analysis the dynamic of this Markov chain is simplified and assumed to be constant.

In this paper, we investigate ESs with weighted recombination on a general convex quadratic function. Since the CMA-ES and most of the recent variants of CMA-ES [3, 23, 25] employ weighted recombination, weighted recombination ESs are among the most important categories of ESs. The first analysis of weighted recombination ESs were done in [4], where the quality gain has been derived on the infinite dimensional sphere function $f : x \mapsto \|x\|^2$. Moreover, the optimal step-size and the optimal recombination weights are derived. The quality gain on a convex quadratic function has been studied in [12,13] for a variant of weighted recombination ESs called $(\mu/\mu_I, \lambda)$-ES that employs the truncation weights, where the weights for μ best candidate solutions are $1/\mu$ and the other weights are zero. We extend and generalize these results with a mathematically rigorous derivation.

The contributions of this paper are as follows.

First, the asymptotic quality gain of the weighted recombination evolution strategy on a convex quadratic function $f(x) = \frac{1}{2}(x - x^*)^{\mathrm{T}}\mathbf{A}(x - x^*)$ with Hessian $\mathbf{A}$ satisfying $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 \ll 1$ is derived for the infinite dimensional search space. The recombination weights optimal for the quality gain turn out to be the same values as the ones derived for the infinite dimensional sphere function [4]. It implies that the optimal weights derived for the infinite dimensional sphere function is optimal independently of the Hessian matrix $\mathbf{A}$ of the objective function and the covariance matrix $\mathbf{C}$ of the sampling distribution. Moreover, we see the dependency of the quality gain on the eigenvalue distribution of the Hessian matrix. It provides a better understanding of the algorithm than our empirical knowledge that the convergence speed of the algorithm with a fixed covariance matrix is roughly proportional to $1/(N \,\mathrm{Cond}(\mathbf{AC}))$ when $\mathbf{C}$ is fixed, whereas our result reveals the dependency of the convergence speed not only on the condition number of $\mathbf{AC}$ but on the eigenvalue distribution of $\mathbf{AC}$ as long as $\mathrm{Tr}((\mathbf{AC})^2) \ll \mathrm{Tr}(\mathbf{AC})^2$. This may lead to a better algorithm design since such an empirical knowledge is used for algorithm design [2].

Second, our proof is rather different from the derivations of the quality gain and the progress rate in previous works. On the one hand results derived in for instance [4,11,16] rely on a geometric intuition of the algorithm in the infinite dimensional search space and on various approximations. On the other hand, the rigorous derivation of the progress rate (or convergence rate of the algorithm with step-size proportional to the optimum) on the sphere function provided for instance in [7, 21] only holds on spherical functions and provides solely a limit without a bound between the finite dimensional convergence rate and its asymptotic limit (see below). In contrast, we provide a novel and mathematically rigorous proof, based on the expression of the expectation of the recombination weights assigned to a given candidate solution as a function of the candidate solution. It is inspired by the previous work [1].

Third, our result is not only asymptotic for N to infinity as opposed to previous rigorous results deriving progress rate [7, 8, 21] (see also [6] for an overview of those results) but we provide an error bound between the finite dimensional quality gain and its limit. The bound shows the dependency of the convergence speed of the quality gain of the weighted recombination ES solving an arbitrary convex quadratic function to its limit on the recombination weights and the eigenvalue distribution of the Hessian matrix of the objective function. Thanks to the explicit bound, we can treat the population size increasing with the dimension of the search space and provide (for instance) a rigorous sufficient condition on the dependency between population size and N such that the per-iteration convergence rate scaling of $O(\lambda/N)$ holds for algorithms with intermediate recombination [11].

This paper is organized as follows. In Section 2, we formally define the evolution strategy with weighted recombination. The quality gain analysis on the infinite dimensional sphere function is revisited. In Section 3, we derive the quality gain bound for a finite dimensional convex quadratic function. The asymptotic quality gain is derived as a consequence. We discuss how the eigenvalue distribution of the Hessian matrix of the objective function influences the quality gain and its convergence speed for $N \to \infty$. In Section 4, we conduct simulations to visualize the effect of different Hessian matrices, the dimension, the recombination weights, the learning rate for the mean vector update, and the step-size. In Section 5, we discuss further topics: the tightness of the derived bound, interpretation of the results for a fixed but non-identity covariance matrix, the dynamics of the mean vector on a convex quadratic function, the geometric interpretation of the optimal setting, the linear convergence proof using the quality gain analysis, and further related works.

[2]The algorithm is "artificial" in that the distance to the optimum is unavailable in practice.

[3]More precisely, $(1 \overset{+}{,} \lambda)$-ES optimizing any function f (that may have more than one global optimum) can not converge towards a given optimum x^* faster in the search space than the artificial algorithm with step-size proportional to the distance to x^*.

2. FORMULATION

2.1 Evolution Strategy with Weighted Recombination

We consider an evolution strategy with weighted recombination. At each iteration $t \geqslant 0$ it samples candidate solutions $X_1, \dots, X_\lambda$ from the N-dimensional normal distribution $\mathcal{N}(\boldsymbol{m}^{(t)}, (\sigma^{(t)})^2\mathbf{I})$, where $\boldsymbol{m}^{(t)} \in \mathbb{R}^N$ is the mean vector and $\sigma^{(t)} > 0$ is the standard deviation, also called the step-size or the mutation strength in the work of Beyer and his co-authors, e.g. [11], and $\mathbf{I}$ is the identity matrix of dimension N. The candidate solutions are evaluated on a given objective function $f : \mathbb{R}^N \to \mathbb{R}$. W.l.g., we assume f to be minimized. Let $i : \lambda$ be the index of the ith best candidate solution among $X_1, \dots, X_\lambda$, i.e., $f(X_{1:\lambda}) \leqslant \dots \leqslant f(X_{\lambda:\lambda})$, and $w_1 \geqslant \cdots \geqslant w_\lambda$ be the weights. W.l.g., we assume $\sum_{i=1}^{\lambda} |w_i| = 1$. Let $\mu_w = 1/\sum_{i=1}^{\lambda} w_i^2$ denote the so-called effective variance selection mass. The mean vector is updated according to

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sum_{i=1}^{\lambda} w_i (X_{i:\lambda} - \boldsymbol{m}^{(t)}) \ , \tag{1}$$

where $c_m > 0$ is the learning rate of the mean vector update, i.e., $\kappa = 1/c_m$ is the ratio between the standard deviation σ and the step-size for the mean update $\sigma \cdot c_m$.

To proceed the analysis with a mathematical rigor, we introduce the weight function as follows

$$W(i; (X_k)_{k=1}^{\lambda}) := \sum_{k=1+l}^{u} \frac{w_k}{u - l}, \tag{2}$$

where $\mathbb{I}\{\text{condition}\}$ is the indicator function which is 1 if the condition is true and 0 otherwise, and l and u are the numbers of strictly better candidate solutions and equally well or better candidate solutions than X_i, respectively, which are defined as follows

$$l = \sum_{j=1}^{\lambda} \mathbb{I}\{f(X_j) < f(X_i)\} \ , \tag{3}$$

$$u = \sum_{j=1}^{\lambda} \mathbb{I}\{f(X_j) \leqslant f(X_i)\} \ . \tag{4}$$

The weight value for X_i is the arithmetic average of the weights w_k for the tie candidate solutions. In other words, all the tie candidate solutions have the same weight values. If there is no tie, the weight value for the ith best candidate solution $X_{i:\lambda}$ is simply w_i. With the weight function, we rewrite the algorithm (1) as follows

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sum_{i=1}^{\lambda} W(i; (X_k)_{k=1}^{\lambda})(X_i - \boldsymbol{m}^{(t)}) \ , \tag{5}$$

or equivalently, letting $Z_k = (X_k - \boldsymbol{m}^{(t)})/\sigma^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sigma^{(t)} \sum_{i=1}^{\lambda} W(i; (\boldsymbol{m}^{(t)} + \sigma^{(t)} Z_k)_{k=1}^{\lambda}) Z_i \ . \tag{6}$$

The above update (5) (or (6)) is equivalent with the original update (1) if there is no tie among λ candidate solutions. If the objective function is a convex quadratic function, there will be no tie with probability one (w.p.1). Therefore, they are equivalent w.p.1.

The motivation of the new formulation is twofold. One is to well define the update even when there is tie, which happens w.p.0, though. The other is a technical reason. In (1) the already sorted candidate solutions $X_{i:\lambda}$ are all correlated and they are not anymore normally distributed. However, they are assumed to be normally distributed in the previous work [4, 12, 13]. To ensure that such an approximation leads to the asymptotically true quality gain limit, a mathematically involved analysis has to be done. See [7, 8, 21] for details. In (5) or (6), the ranking computation is a part of the weight function and X_i are still independent and normally distributed. This allows us in a rigorous and novel approach to derive the quality gain on a convex quadratic function.

In the analysis of this paper, we do not consider the adaptation of step-size $\sigma^{(t)}$. Instead we investigate the response of the progress measure to the input step-size $\sigma^{(t)} = \sigma$.

2.2 Quality Gain Analysis on a Spherical Function

In general, it is too difficult to analyze the Markov chain defined by rank-based stochastic algorithms on the continuous domain. One won't obtain the explicit formula of the convergence rate of an algorithm. Instead, we study the expected amount of the progress in one algorithmic iteration. The quality gain [10, 24] is one of a common measure to evaluate such progress. It is defined as the expectation of the relative decrease of the function value. In this paper we define the quality gain as the conditional expectation of the relative decrease of the function value given the mean vector $\boldsymbol{m}^{(t)} = \boldsymbol{m}$ and the step-size $\sigma^{(t)} = \sigma$ as

$$\phi(\boldsymbol{m}, \sigma) = \frac{\mathbb{E}[f(\boldsymbol{m}^{(t)}) - f(\boldsymbol{m}^{(t+1)}) \mid \boldsymbol{m}^{(t)} = \boldsymbol{m}, \sigma^{(t)} = \sigma]}{f(\boldsymbol{m}^{(t)}) - f(x^*)} \ , \tag{7}$$

where $x^* \in \mathbb{R}^N$ is (one of) the global minimum of f. Note that the quality gain depends also on the weights $(w_k)_{k=1}^{\lambda}$, the learning rate c_m, and the dimension N. To avoid division by zero, we assume that $\boldsymbol{m} \in \mathbb{R}^N \setminus \{x^*\}$.

In [4], the algorithm (1) solving a spherical function $f(x) = \|x\|^2$ is analyzed. For this purpose, the *normalized step-size* and the *normalized quality gain* are introduced. The normalized step-size at $\boldsymbol{m}$ is defined as

$$\bar{\sigma} = \frac{\sigma c_m N}{\|\boldsymbol{m}\|} \ , \tag{8}$$

and the normalized quality gain is defined as the quality gain ϕ scaled by $N/2$, namely,

$$\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = \frac{N}{2} \phi(\boldsymbol{m}, \sigma = \bar{\sigma}\|\boldsymbol{m}\|/(c_m N)) \ . \tag{9}$$

It is stated that by taking $N \to \infty$, the normalized quality gain converges to the *asymptotic normalized quality gain*

$$\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda}) = -\bar{\sigma} \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] - \frac{\bar{\sigma}^2}{2} \sum_{i=1}^{\lambda} w_i^2 \ , \tag{10}$$

where $\mathbb{E}[\mathcal{N}_{i:\lambda}]$ is the expected value of the ith smallest order statistics $\mathcal{N}_{i:\lambda}$ from λ independent populations sampled from the standard normal distribution (a formal proof of this result is presented in [7] with the detailed proof for the uniform integrability done in [21]). For a sufficiently large N, one can approximate the quality gain as $\phi(\boldsymbol{m}, \sigma) \approx (N/2)\bar{\phi}_\infty(\bar{\sigma} = \sigma c_m N/\|\boldsymbol{m}\|, (w_k)_{k=1}^{\lambda})$.

Consider the optimal parameter setting that maximize $\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda})$ in (10). Remember $\sum_{i=1}^{\lambda} |w_i| = 1$ and $\mu_w = 1/\sum_{i=1}^{\lambda} w_i^2$. The optimal weights w_i^* are

$$w_i^* = -\frac{\mathbb{E}[\mathcal{N}_{i:\lambda}]}{\sum_{i=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]|} \ . \tag{11}$$

Moreover, given $(w_k)_{k=1}^{\lambda}$, the optimal normalized step-size $\bar{\sigma}^*$ is

$$\bar{\sigma}^*((w_k)_{k=1}^{\lambda}) = -\mu_w \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] \ . \tag{12}$$

The optimal value of the asymptotic normalized quality gain is $\bar{\phi}_{\infty}(\bar{\sigma}^*((w_k^*)_{k=1}^{\lambda}), (w_k^*)_{k=1}^{\lambda}) = \frac{1}{2}\sum_{i=1}^{\lambda} \mathbb{E}[\mathcal{N}_{i:\lambda}]^2$, which is roughly estimated by $\lambda/2$ if λ is sufficiently large, say $\lambda \geqslant 10^2$. Therefore, the quality gain is approximated by λ/N.

We have three remarks on the optimal normalized quality gain. First, it tells that the quality gain is proportional to $1/N$. Note that we know from [26] that a comparison-based algorithm can not achieve the convergence rate better than $1/N$. One can not expect a faster convergence except a constant factor. Second, the quality gain is to measure the improvement in one iteration. If we generate and evaluate λ candidate solutions every iteration, the quality gain per evaluation (f-call) is $1/\lambda$ times smaller, i.e., the quality gain per evaluation is $1/N$, rather than λ/N. It implies that the number of iterations to achieve the same amount of the quality gain is inversely proportional to λ, which is ideal when the algorithm is implemented on a parallel computer. The third remark is that the result is obtained in the limit $N \to \infty$ while λ is fixed. Therefore, it is implicitly assumed that $\lambda \ll N$. The second remark above holds only when $\lambda \ll N$. In practice, the quality gain per iteration tends to level out as λ increases.

Figure 1 shows how the asymptotic normalized quality gain divided by λ scales with λ when the optimal step-size $\sigma = \|\boldsymbol{m}\|\bar{\sigma}^*((w_k)_{k=1}^{\lambda})/(c_m N)$ is employed. Four different weights schemes are employed, the optimal weights ($w_k \propto -\mathbb{E}[\mathcal{N}_{k:\lambda}]$), the weights used in the CMA-ES ($w_k \propto \max(\ln\left(\frac{\lambda+1}{2}\right) - \ln(k), 0)$), and the truncation weights ($w_k = 1/\mu$ for $k = 1, \dots, \mu$ and $w_k = 0$ for $k = \mu+1, \dots, \lambda$) with $\mu = \lfloor \lambda/4 \rfloor$ and $\mu = \lfloor \lambda/10 \rfloor$. All the weights are scaled so that $\sum_{k=1}^{\lambda} |w_k| = 1$. The expected value of the normal order statistics $\mathcal{N}_{i:\lambda}$ for $i \leqslant \lceil \lambda/2 \rceil$ is approximated by the Blom's formula

$$\mathbb{E}[\mathcal{N}_{i:\lambda}] \approx \Phi^{-1}\left(\frac{i-\alpha}{\lambda - 2\alpha + 1}\right), \quad \alpha = 0.375 \ , \tag{13}$$

where Φ is the cumulative density function of the standard normal distribution. Note that $\mathbb{E}[\mathcal{N}_{1:\lambda}] < \dots < \mathbb{E}[\mathcal{N}_{\lambda:\lambda}]$ and $\mathbb{E}[\mathcal{N}_{i:\lambda}] = -\mathbb{E}[\mathcal{N}_{\lambda-i:\lambda}]$. When the optimal weights are used, it goes up to 0.5 as λ increases (implying $\sum_{k=1}^{\lambda} |\mathbb{E}[\mathcal{N}_{i:\lambda}]| \approx \lambda$ for a sufficiently large λ). On the other hand, the quality gain with only nonnegative weights can not be above 0.25, which will be obtained if the first half of the weights are proportional to the optimal setting and the last half of the weights are zero. The CMA type weights well approximate the optimal nonnegative weights.

The optimal step-size (12) depends on $(w_k)_{k=1}^{\lambda}$, in particular, depends on μ. For example, consider the truncation weights with a fixed ratio λ/μ. Then, we find $\bar{\sigma}^* = \sum_{i=1}^{\mu}(-\mathbb{E}[\mathcal{N}_{i:\lambda}]) \in \Theta(\mu)$. It implies that the optimal step-size is roughly proportional to μ. However, since the optimality holds for $N \to \infty$ while λ is fixed, it is implicitly assumed that $\lambda \ll N$. Therefore, $\bar{\sigma}^* \in \Theta(\mu)$ may not hold if $\lambda \not\ll N$.

The asymptotic normalized quality gain (10) can be seen as a quadratic function of $\bar{\sigma}$. Figure 2 shows how the normalized quality gain divided by λ scales with $\bar{\sigma}/\bar{\sigma}^*$. The maximum normalized quality gain is reached when $\bar{\sigma}/\bar{\sigma}^* = 1$. With a smaller step-size, we observe a slower convergence.

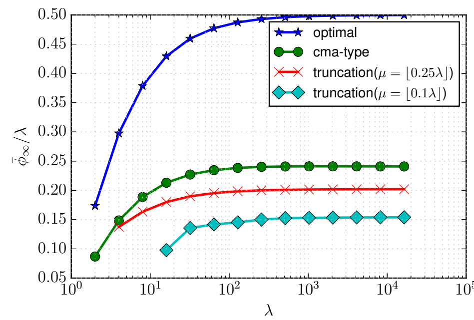


Figure 1: The asymptotic normalized quality gain $\bar{\phi}_{\infty}(\bar{\sigma}^*((w_k)_{k=1}^{\lambda}), (w_k)_{k=1}^{\lambda})$ divided by λ vs λ.

On the other hand, we observe a rapid drop of the normalized quality gain if we increase the step-size. With $\bar{\sigma}/\bar{\sigma}^* = 2$, we see no progress. With a larger step-size, we observe the negative progress, i.e., the divergence of the mean vector.

The normalized quality gain depends only on the normalized step-size $\bar{\sigma}$ and the weights $(w_k)_{k=1}^{\lambda}$. Since the normalized step-size does not change if we multiply c_m by some factor and divide σ by some factor, one can say that c_m doesn't have any impact on the asymptotic normalized quality gain, hence on the quality gain, as long as σc_m is constant. This is unintuitive and not true in a finite dimensional space. The step-size σ realizes the standard deviation of the sampling distribution and it has an impact on the ranking of the candidate solutions. On the other hand, the product σc_m is the step-size of the $\boldsymbol{m}$-update, which must depend on the ranking of the candidate solutions. The asymptotic normalized quality gain provided above tells us that the ranking of the candidate solutions is independent of $\bar{\sigma}$ in the infinite dimensional space.

3. QUALITY GAIN ANALYSIS ON A GENERAL QUADRATIC FUNCTION

In this section we derive the normalized quality gain bound and the asymptotic normalized quality gain of the weighted recombination ES (1) minimizing a quadratic function with its Hessian $\nabla\nabla f(x) = \mathbf{A}$ assumed to be nonnegative definite and symmetric, i.e.,

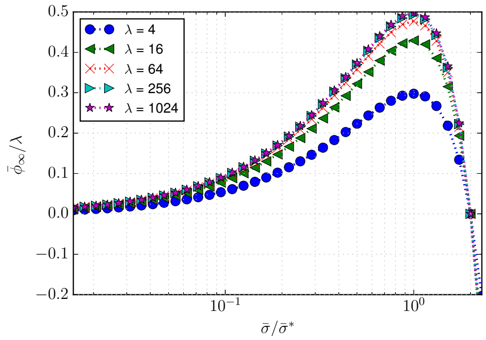
$$f(x) = \frac{1}{2}(x - x^*)^{\mathrm{T}} \mathbf{A} (x - x^*) \ , \tag{14}$$

where $x^* \in \mathbb{R}^N$ is the global optimal solution[4]. The performance of an algorithm on a quadratic function is essential to understand the local behavior of the algorithm since an arbitrary twice continuously differentiable function can be approximated by a quadratic function locally.

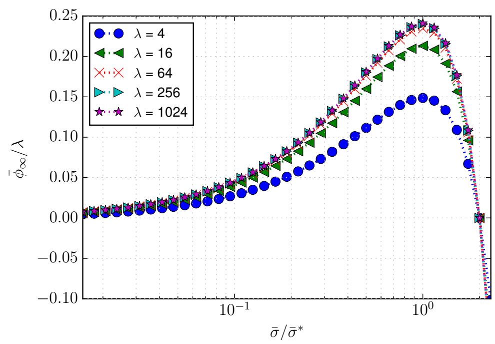
3.1 Normalized Quality Gain and Normalized Step-Size

We introduce the normalized step-size and the normalized quality gain as we did in (8) and (9) for the sphere function. First of all, if the objective function is homogeneous around the optimal solution x^*, the optimal step-size must be a

[4]A nonnegative definite matrix $\mathbf{A}$ is a matrix having only nonnegative eigenvalues, i.e., $x^{\mathrm{T}}\mathbf{A}x \geqslant 0$ for all $x \in \mathbb{R}^N$. If $\mathbf{A}$ is not full rank, i.e., it has a zero eigenvalue, the optimum x^* is not unique.



(a) Optimal Weights



(b) CMA Weights

Figure 2: The asymptotic normalized quality gain divided by λ vs the normalized step-size ratio $\bar{\sigma}/\bar{\sigma}^*$. The results for the population size $\lambda = 4^i$ for $i = 1, 2, 3, 4, 5$ are shown.

homogeneous function of degree one with respect to $\boldsymbol{m} - x^*$. This is formally stated in the following proposition. The proof is found in Appendix A.

PROPOSITION 3.1. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a homogeneous function of degree n, i.e., $f(\alpha \cdot x) = \alpha^n f(x)$ for a fixed integer $n > 0$ for any $\alpha > 0$ and any $x \in \mathbb{R}^N$. Consider the weighted recombination ES (1) minimizing a function $g : x \mapsto f(x - x^*)$. Then, the quality gain is scale-invariant, i.e., $\phi(x^* + (\boldsymbol{m} - x^*), \sigma) = \phi(x^* + \alpha(\boldsymbol{m} - x^*), \alpha\sigma)$ for any $\alpha > 0$. Moreover, the optimal step-size $\sigma^* = \operatorname{argmax}_{\sigma \geqslant 0} \phi(\boldsymbol{m}, \sigma)$, if it is well-defined, is a function of $\boldsymbol{m} - x^*$. For the sake of simplicity we write the optimal step-size as a map $\sigma^* : \boldsymbol{m} - x^* \mapsto \sigma^*(\boldsymbol{m} - x^*)$. It is a homogeneous function of degree 1, i.e., $\sigma^*(\alpha \cdot (\boldsymbol{m} - x^*)) = \alpha\sigma^*(\boldsymbol{m} - x^*)$ for any $\alpha > 0$.*

Note that the function $\boldsymbol{m} \mapsto \|\nabla f(\boldsymbol{m})\| = \|\mathbf{A}(\boldsymbol{m} - x^*)\|$ is homogeneous of degree one around x^*. However, this function is not invariant to scaling of a function, i.e., scaling of Hessian $\mathbf{A}$. To make it scale invariant, we will divide it by the trace $\mathrm{Tr}(\mathbf{A})$ of Hessian $\mathbf{A}$. Therefore, the function $\boldsymbol{m} \mapsto \|\nabla f(\boldsymbol{m})\| / \mathrm{Tr}(\mathbf{A})$ is our candidate for the optimal step-size. We define the normalized step-size and the scale-invariant step-size for a quadratic function as follows.

DEFINITION 3.2. *For a quadratic function (14), the* normalized step-size $\bar{\sigma} > 0$ *is defined as*

$$\bar{\sigma} = \frac{\sigma c_m \mathrm{Tr}(\mathbf{A})}{\|\nabla f(\boldsymbol{m})\|} \quad . \tag{15}$$

In other words, the step-size is given by

$$\sigma = \frac{\bar{\sigma}\|\nabla f(\boldsymbol{m})\|}{c_m \mathrm{Tr}(\mathbf{A})} \quad . \tag{16}$$

We call it the scale-invariant step-size *for a quadratic function (14).*

The normalized quality gain of weighted recombination ES with scale-invariant step-size on a quadratic function is then defined as follows.

DEFINITION 3.3. *Let $g : \mathbb{R}^N \to \mathbb{R}$ be the $\boldsymbol{m}$-dependent scaling factor of the normalized quality gain defined by*

$$g(\boldsymbol{m}) = \frac{\|\nabla f(\boldsymbol{m})\|^2}{f(\boldsymbol{m}) \mathrm{Tr}(\mathbf{A})} \quad . \tag{17}$$

The normalized quality gain *for a quadratic function is defined as*

$$\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = \frac{\phi(\boldsymbol{m}, \sigma = \bar{\sigma}\|\nabla f(\boldsymbol{m})\| / (c_m \mathrm{Tr}(\mathbf{A})))}{g(\boldsymbol{m})} \quad . \tag{18}$$

In other words,

$$\phi\left(\boldsymbol{m}, \sigma = \bar{\sigma}\|\nabla f(\boldsymbol{m})\| / (c_m \mathrm{Tr}(\mathbf{A}))\right) = g(\boldsymbol{m})\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) \quad . \tag{19}$$

Note that the normalized step-size and the normalized quality gain defined above agree with (8) and (9), respectively, if $f(x) = \|x\|^2$, where $\mathbf{A} = 2\mathbf{I}$, $\nabla f(\boldsymbol{m}) = \boldsymbol{m}$ and $g(\boldsymbol{m}) = 2/N$. Moreover, they are equivalent to Eq. (4.104) in [11] introduced to analyze the $(1+\lambda)$-ES and the $(1, \lambda)$-ES. The same normalized step-size has been used for $(\mu/\mu_I, \lambda)$-ES [12, 13]. See Section 4.3.1 of [11] for the motivation of these normalization.

3.2 Theorem: Normalized Quality Gain on Convex Quadratic Functions

The following theorem provides an upper bound of the difference between the normalized quality gain on a finite dimensional convex quadratic function and the asymptotic normalized quality gain derived on the sphere function.

THEOREM 3.4 (NORMALIZED QUALITY GAIN BOUND). *Consider the weighted recombination evolution strategy (6) solving a convex quadratic function (14). Let the normalized step-size and the normalized quality gain defined as (15) and (18). Then,*

$$\begin{aligned}
&\sup_{\boldsymbol{m} \in \mathbb{R}^N \setminus \{x^*\}} \left| \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda) \right| \\
&\leqslant \frac{3}{(4\pi)^{\frac{1}{3}}} \frac{\bar{\sigma}^{\frac{5}{3}}}{c_m^{\frac{2}{3}}} \lambda(\lambda - 1) \max_{k \in [\![1, \lambda - 1]\!]} |w_{k+1} - w_k| \left[\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right]^{\frac{1}{3}} \\
&\quad + \frac{\bar{\sigma}^2}{2^{\frac{1}{2}}} \left[\left[(\lambda^2 - \lambda) \sum_{k=1}^{\lambda} \sum_{l=k+1}^{\lambda} w_k^2 w_l^2 \right]^{\frac{1}{2}} \right. \\
&\qquad \left. + \left[\lambda \sum_{k=1}^{\lambda} w_k^4 \right]^{\frac{1}{2}} \right] \left[\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right]^{\frac{1}{2}},
\end{aligned} \tag{20}$$

where $\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda)$ is the function defined in (10).

The first consequence of Theorem 3.4 is the generalization of the infinite dimensional analysis on the sphere function in [4] to a convex quadratic function. Let $(\mathbf{A}_N)_{N \in \mathbb{N}}$ be the sequence of Hessian matrices $\mathbf{A}_N$ of a convex quadratic function (14) of dimension $N \in \mathbb{N}_+$. Under the condition

$\mathrm{Tr}(\mathbf{A}_N^2)/\mathrm{Tr}(\mathbf{A}_N)^2 \to 0$ as $N \to \infty$, we can prove that the normalized quality gain defined in (18) converges to the unique limit $\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda)$. In particular, the limit agrees with the one of (10) derived for the sphere function. The following corollary formalizes it, which can be immediately derived from Theorem 3.4.

Corollary 3.5 (Normalized Quality Gain Limit). *Suppose that* $\mathrm{Tr}(\mathbf{A}_N^2)/\mathrm{Tr}(\mathbf{A}_N)^2 \to 0$ *as* $N \to \infty$. *Then,*

$$\lim_{N\to\infty} \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \left|\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda)\right| = 0 \ . \tag{21}$$

It tells that the quality gain on a convex quadratic function with Hessian $\mathbf{A}$ such that $\mathrm{Tr}(\mathbf{A}^2) \ll \mathrm{Tr}(\mathbf{A})^2$ under the scale-invariant step-size (Definition 3.2) is approximated by the product of the normalization factor $g(\boldsymbol{m})$ and the asymptotical normalized quality gain as

$$\begin{aligned}\phi(\boldsymbol{m}, \sigma) &= g(\boldsymbol{m})\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) \approx g(\boldsymbol{m})\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda) \\ &= \frac{\|\nabla f(\boldsymbol{m})\|^2}{f(\boldsymbol{m})\,\mathrm{Tr}(\mathbf{A})}\left[-\bar{\sigma}\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] - \frac{\bar{\sigma}^2}{2}\sum_{i=1}^{\lambda} w_i^2\right] \ .\end{aligned} \tag{22}$$

Asymptotically, the quality gain can be decomposed in two parts: the normalization factor $g(\boldsymbol{m})$ that depends on the mean vector and the eigenvalue distribution of $\mathbf{A}$, and the asymptotic normalized quality gain that depends on the normalized step-size and the recombination weights. Note that the RHS of (22) agrees with the limit of the one derived for the $(1, \lambda)$-ES (Eq. (4.108) in [11]) if $c_m = 1$, $w_1 = 1$ and $w_i = 0$ for $i \geqslant 2$[5]. Moreover, it coincides with the limit of the quality gain for $(\mu/\mu_I, \lambda)$-ES deduced from the results obtained in [12, 13] if $c_m = 1$, $w_i = 1/\mu$ for $i = 1, \dots, \mu$ and $w_i = 0$ for $i = \mu + 1, \dots, \lambda$.

The second consequence is a further generalization of Corollary 3.5 that allows us to take λ increasing to $+\infty$ as $N \to \infty$. It reflects the practical setting that we often set λ dependently on N. The optimal normalized step-size $\bar{\sigma}^*$ (with a slight abuse of notation) given in (12) depends on $(w_k)_{k=1}^\lambda$ and typically scales up in the order of $O(\lambda)$. Moreover, if $\bar{\sigma} = \beta\bar{\sigma}^*$ for some $\beta \in (0, 2)$, the asymptotic normalized quality gain is

$$\begin{aligned}\bar{\phi}_\infty(\beta\bar{\sigma}^*, (w_k)_{k=1}^\lambda) &= (\beta - \beta^2/2)(\bar{\sigma}^*)^2 \textstyle\sum_{i=1}^\lambda w_i^2 \\ &= (\beta - \beta^2/2)\mu_w(\textstyle\sum_{i=1}^\lambda w_i \mathbb{E}[\mathcal{N}_{i:\lambda}])^2,\end{aligned}$$

where the right-most side is maximized when $w_i \propto -\mathbb{E}[\mathcal{N}_{i:\lambda}]$ and the maximum value is $(\beta - \beta^2/2)\sum_{i=1}^\lambda |\mathbb{E}[\mathcal{N}_{i:\lambda}]| \in \Theta(\lambda)$ (see (25) below). Since both the normalized step-size and the normalized quality gain typically diverges towards infinity, the convergence of the normalized quality gain pointwise with respect to $\bar{\sigma}$ such as (21) is not sufficient. In the following, we provide a sufficient condition to show the convergence of the error rate between the normalized quality gain and the asymptotic normalized quality gain uniformly for $\bar{\sigma} \in [\epsilon\bar{\sigma}^*, (2 - \epsilon)\bar{\sigma}^*]$.

[5] In [11], the author claims that the condition on $\mathbf{A}$ to obtain the limit (22) for $N \to \infty$ is $\bar{\sigma}\,\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \ll \mathrm{Tr}(\mathbf{A})$ (Eq. (4.107) in [11], called the Sphere model condition). It can be satisfied if $\bar{\sigma} \ll 1$ but $\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \approx \mathrm{Tr}(\mathbf{A})$, however, $d_1(\mathbf{A}) \ll \mathrm{Tr}(\mathbf{A})$ is assumed to derive Eq. (4.108) and is not satisfied when $\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \approx \mathrm{Tr}(\mathbf{A})$. In Theorem 3.4, if we have $\bar{\sigma} \to \infty$ but $\mathrm{Tr}(\mathbf{A}^2)^{\frac{1}{2}} \not\ll \mathrm{Tr}(\mathbf{A})$, both LHS and RHS of (20) converge to zero, but the RHS converges faster.

Corollary 3.6. *Let* λ_N *be the population size for the dimension* N *and* $(w_k^{(N)})_{k=1}^{\lambda_N}$ *be the sequence of weights of length* λ_N. *Let* $\bar{\sigma}_N^*$ *be the optimal normalized step-size defined in* (12) *given* $(w_k^{(N)})_{k=1}^{\lambda_N}$. *Suppose that*

$$\max\left[\lambda^2, \frac{\lambda^6\mu_w^2 \max_k |w_{k+1} - w_k|^3}{-\sum_{i=1}^\lambda w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]}\right] \in o\left[\frac{\mathrm{Tr}(\mathbf{A}_N)^2}{\mathrm{Tr}(\mathbf{A}_N^2)}\right] \ , \tag{23}$$

where $\max_k$ *is taken over* $k \in [\![1, \lambda - 1]\!]$. *Then, for any* $\epsilon \in (0, 1)$,

$$\lim_{N\to\infty} \sup_{\bar{\sigma}} \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \frac{\left|\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \bar{\phi}_\infty(\bar{\sigma}, (w_k^{(N)})_{k=1}^{\lambda_N})\right|}{\bar{\phi}_\infty(\bar{\sigma}, (w_k^{(N)})_{k=1}^{\lambda_N})} = 0 \ , \tag{24}$$

where $\sup_{\bar{\sigma}}$ *is taken over* $\bar{\sigma} \in [\epsilon\bar{\sigma}_N^*, (2 - \epsilon)\bar{\sigma}_N^*]$.

Consider, for example, the truncation weights with the number of parents μ_N and the sphere function ($\mathbf{A} = 2\mathbf{I}$). Then, the condition reads $\lambda_N^6/(-\sum_{i=1}^{\mu_N}\mathbb{E}[\mathcal{N}_{i:\lambda}]) \in o(N)$. In particular, if λ_N/μ_N is constant (or upper bounded), it reads $\lambda_N^5 \in o(N)$. If μ_N is upper bounded, then the condition reads $\lambda_N^6 \in o(N)$. If we consider the optimal weights (11), using the trivial inequality $|w_{k+1} - w_k| \leqslant |w_\lambda - w_1|$ and the fact (see for example Example 8.1.1 in [14]) that

$$\begin{aligned}&\lim_{\lambda\to\infty} \frac{\mathbb{E}[\mathcal{N}_{\lambda:\lambda}] - \mathbb{E}[\mathcal{N}_{1:\lambda}]}{2(2\ln(\lambda))^{\frac{1}{2}}} = 1 \ , \\ &\lim_{\lambda\to\infty} \frac{1}{\lambda}\sum_{i=1}^\lambda \mathbb{E}[\mathcal{N}_{i:\lambda}]^2 = 1 \ , \\ &\lim_{\lambda\to\infty} \frac{1}{\lambda}\sum_{i=1}^\lambda |\mathbb{E}[\mathcal{N}_{i:\lambda}]| = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \ ,\end{aligned} \tag{25}$$

the condition reads $\lambda_N^5(\ln(\lambda_N))^{\frac{3}{2}} \in o(N)$. In the state-of-the-art algorithm, namely the CMA-ES, the default population size is $\lambda_N = 4 + \lfloor 3\ln(N)\rfloor$ and the weights are similar to the positive part of the optimal weights and the others are zero. This case fits in Corollary 3.6 as well.

This condition $\lambda_N \in o(N^{1/6})$ is a rigorous (but seemingly not tight) bound for the scaling of λ such that the per-iteration convergence rate of a $(\mu/\mu, \lambda)$-ES with a fixed λ/μ on the sphere function scales like $O(\lambda/N)$ [11, Equation 6.140].

Remark that contrary to the spherical case where we can show that for any dimension, the algorithm with scale-invariant step-size $\sigma = \sigma^{\mathrm{opt}}\|x - x^*\|$ (for a proper choice of constant σ^{opt} and optimal weights) is optimal [20, 21], we cannot prove here the optimality of the scale-invariant step-size (16) for a finite dimension. However for a quadratic function with $\mathrm{Tr}(\mathbf{A}^2) \ll \mathrm{Tr}(\mathbf{A})^2$, the algorithm with the scale-invariant step-size (16) achieves the quality gain greater than $1 - \epsilon$ times the optimal quality gain $g(\boldsymbol{m})\bar{\phi}_\infty$, where $\epsilon > 0$ is the right-hand side of (20) divided by $\bar{\phi}_\infty$.

3.3 Effect of the Eigenvalue Distribution of the Hessian Matrix

The theorem tells that the optimal weights are the same on any infinite dimensional quadratic function satisfying the condition $\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2 \to 0$. In particular, the optimal weights on such quadratic functions are the same as the optimal weights on the infinite dimensional sphere function. It is a nice feature since we do not need to tune the weight values depending on the function.

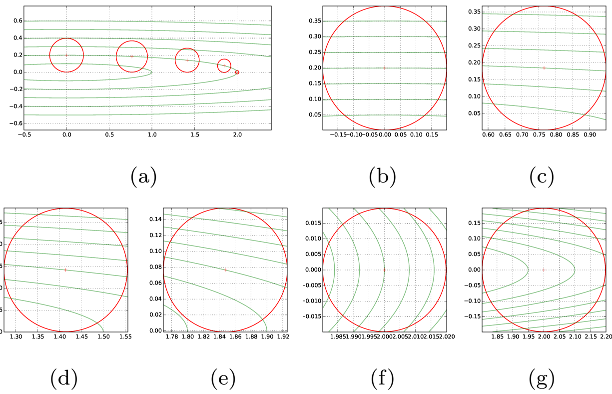


Figure 3: The scale-invariant step-size on $f(x) = x^{\mathrm{T}}\mathbf{A}x/2$ with $\mathbf{A} = \operatorname{diag}(1, 100)$. (a) The circles with radius $\|\nabla f(\boldsymbol{m})\|/\operatorname{Tr}(\mathbf{A})$ centered at $\boldsymbol{m} = 2\mathbf{A}^{-\frac{1}{2}}[\cos(\theta), \sin(\theta)]^{\mathrm{T}}$ with $\theta = \frac{1}{2}\pi, \frac{3}{8}\pi, \frac{1}{4}\pi, \frac{1}{8}\pi, 0$. (b–f) The contour lines focused in each circle. (g) The contour lines focused in the circle with the same radius as (b) centered at the same point as (f).

The optimal normalized step-size and the optimal normalized quality gain are independent of $\mathbf{A}$. However, the step-size and the quality gain depends on it. When the weights and the normalized step-size are fixed, the step-size and the quality gain are

$$\sigma = \frac{\bar{\sigma}}{c_m}\frac{\|\nabla f(\boldsymbol{m})\|}{\operatorname{Tr}(\mathbf{A})}, \quad \phi(\boldsymbol{m}, \sigma) = g(\boldsymbol{m})\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^{\lambda}) \ . \tag{26}$$

Since $\nabla f(\boldsymbol{m}) = \mathbf{A}(\boldsymbol{m} - x^*)$, we have

$$\frac{d_N(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})}\|\boldsymbol{m} - x^*\| \leqslant \frac{\|\nabla f(\boldsymbol{m})\|}{\operatorname{Tr}(\mathbf{A})} \leqslant \frac{d_1(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})}\|\boldsymbol{m} - x^*\|$$
$$\frac{d_N(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})} \leqslant \frac{g(\boldsymbol{m})}{2} \leqslant \frac{d_1(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})} \ ,$$

where $d_i(\mathbf{A})$ are the ith greatest eigenvalue of $\mathbf{A}$. The lower and upper equalities for both of the above inequalities hold if and only if $\boldsymbol{m} - x^*$ is parallel to the eigenspace corresponding to the smallest and largest eigenvalues of $\mathbf{A}$, respectively.

On the surface of the hyper ellipsoid centered at the optimum x^*, the optimal step-size can be different by the factor of at most $\operatorname{Cond}(\mathbf{A}) = d_1(\mathbf{A})/d_N(\mathbf{A})$. This is visualized in Figure 3. Each circle corresponds to the equal density line of the normal distribution with the mean $\boldsymbol{m}$ and the standard deviation $\|\nabla f(\boldsymbol{m})\|/\operatorname{Tr}(\mathbf{A})$. If we focus on the area around each circle, which is the right area to look at since the candidate solutions are produced around there, the function landscape looks like a parabolic ridge function.

In the above discussion we assumed that $\mathbf{A}$ is positive definite, i.e., $d_1(\mathbf{A}) \geqslant \ldots \geqslant d_N(\mathbf{A}) > 0$. However, the condition $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2 \to 0$ can be met even if $\mathbf{A}$ is positive semidefinite. Let M be the mathematical rank of $\mathbf{A}$. That is, $d_1(\mathbf{A}) \geqslant \ldots \geqslant d_M(\mathbf{A}) > 0$ and $d_{M+1}(\mathbf{A}) = \cdots = d_N(\mathbf{A})$. In this case, the kernel of $\mathbf{A}$ (the eigenspace corresponding to zero eigenvalue) does not affect the objective function value. The condition $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2 \to 0$ requires the dimension M of the effective search space to tend to the infinity as $N \to \infty$. Let $\boldsymbol{m}^+$ and $\boldsymbol{m}^-$ be the decomposition of $\boldsymbol{m}$ such that $\boldsymbol{m}^-$ is the projection of $\boldsymbol{m}$ onto the subspace through x^* spanned by the eigenvectors of $\mathbf{A}$ corresponding to the zero eigenvalue, and $\boldsymbol{m}^+ = \boldsymbol{m} - \boldsymbol{m}^-$. Then, we have

$$\frac{d_M(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})} \leqslant \frac{g(\boldsymbol{m})}{2} \leqslant \frac{d_1(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})}$$
$$\frac{d_M(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})}\|\boldsymbol{m}^+\| \leqslant \frac{\|\nabla f(\boldsymbol{m})\|}{\operatorname{Tr}(\mathbf{A})} \leqslant \frac{d_1(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})}\|\boldsymbol{m}^+\| \ .$$

In this case, $g(\boldsymbol{m})$ can be $2/M$ if $d_1(\mathbf{A}) = \cdots = d_M(\mathbf{A}) > 0$ and $d_i(\mathbf{A}) = 0$ for $i \in [\![M+1, N]\!]$. The quality gain is then proportional to $2/M$, instead of $2/N$. That is, the optimal evolution strategy (evolution strategy with the optimal step-size) solves the quadratic function with the effective rank M defined on the N dimensional search space as efficiently as it solves its projection onto the effective search space (hence n dimensional search space).

Table 1 summarizes $d_N(\mathbf{A})/\operatorname{Tr}(\mathbf{A})$ and $d_1(\mathbf{A})/\operatorname{Tr}(\mathbf{A})$, which are the lower and upper bound of $g(\boldsymbol{m})/2$, and $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2$ for different types of $\mathbf{A}$. Note that the lower bound and the upper bound of $g(\boldsymbol{m})$, i.e., the worst and the best possible quality gain coefficients, are $d_N(\mathbf{A})/\operatorname{Tr}(\mathbf{A})$ and $d_1(\mathbf{A})/\operatorname{Tr}(\mathbf{A})$, respectively. The value of the fraction $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2$ tells us how fast the normalized quality gain converges as $N \to \infty$. If the condition number $\alpha = \operatorname{Cond}(\mathbf{A})$ is fixed, the worst case is maximized when the function has a discus type structure ($d_1(\mathbf{A}) = \alpha$ and $d_2(\mathbf{A}) = \cdots = d_N(\mathbf{A}) = 1$), and is minimized when the function has a cigar type structure ($d_1(\mathbf{A}) = \cdots = d_{N-1}(\mathbf{A}) = \alpha$ and $d_N(\mathbf{A}) = 1$). The value of $d_N(\mathbf{A})/\operatorname{Tr}(\mathbf{A})$ will be close to $1/N$ as $N \to \infty$ for the discus type function, whereas it will be close to $1/(N\alpha)$ for the cigar. Therefore, if $N \gg \alpha$, the worst case quality gain on the discus type function is as high as the quality gain on the spherical function, whereas the worst case quality gain on the cigar function is $1/\alpha$ times smaller. On the other hand, the inequality $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2 < 1/(N-1)$ on the cigar type function holds independently of α, while $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2$ depends heavily on α on the discus type function. The fraction will not be sufficiently small and we can not approximate the normalized quality gain by Theorem 3.4 unless $\alpha \ll N$. Therefore, the theorem is not appropriate for functions of discus type if the condition number is higher than the dimension of interest[6].

3.4 Proof of the Theorem

Here we sketch the main line of the proof of Theorem 3.4. First we expand the normalized quality gain. Let

$$\mathbf{e} = \nabla f(\boldsymbol{m})/\|\nabla f(\boldsymbol{m})\| \ . \tag{27}$$

With (15), it is easy to see that

$$f(\boldsymbol{m} + \sigma z) - f(\boldsymbol{m}) = \frac{\|\nabla f(\boldsymbol{m})\|^2}{\operatorname{Tr}(\mathbf{A})}\left(\frac{\bar{\sigma}}{c_m}\mathbf{e}^{\mathrm{T}}z + \frac{1}{2}\left(\frac{\bar{\sigma}}{c_m}\right)^2\frac{z^{\mathrm{T}}\mathbf{A}z}{\operatorname{Tr}(\mathbf{A})}\right) \ . \tag{28}$$

Using (28) with $z = (\boldsymbol{m}^{(t+1)} - \boldsymbol{m}^{(t)})/\sigma$, the normalized quality gain on a convex quadratic function can be written as

$$\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) = -\left(\bar{\sigma}\sum_{i=1}^{\lambda}\mathbb{E}[W(i; (X_k)_{k=1}^{\lambda})\mathbf{e}^{\mathrm{T}}Z_i] + \frac{\bar{\sigma}^2}{2}\sum_{i=1}^{\lambda}\sum_{j=1}^{\lambda}\mathbb{E}\left[W(i; (X_k)_{k=1}^{\lambda})W(j; (X_k)_{k=1}^{\lambda})\frac{Z_i^{\mathrm{T}}\mathbf{A}Z_j}{\operatorname{Tr}(\mathbf{A})}\right]\right) \ , \tag{29}$$

[6] However, the worst case scenario on the discus type function, $1/(\alpha + (N-1))$, describes an empirical observation that the convergence speed of evolution strategy with isotropic distribution does not scale up with N for $N \ll \alpha$.

Table 1: Different types of the eigenvalue distributions of $\mathbf{A}$. The second to fourth types (discus: $d_1(\mathbf{A}) = \alpha$ and $d_2(\mathbf{A}) = \cdots = d_N(\mathbf{A}) = 1$, cigar: $d_1(\mathbf{A}) = \cdots = d_{N-1}(\mathbf{A}) = \alpha$ and $d_N(\mathbf{A}) = 1$) have the condition number $\mathrm{Cond}(\mathbf{A}) = d_1(\mathbf{A})/d_N(\mathbf{A}) = \alpha$, while the last type has the condition number αN. The third and the fifth types has the eigenvalues $d_i(\mathbf{A}) = \alpha^{\frac{i-1}{N-1}}$ and $d_i(\mathbf{A}) = \alpha \cdot i$, respectively.

Type	$\frac{d_N(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})}$	$\frac{d_1(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})}$	$\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}$
Sphere	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$
Discus	$\frac{1}{(N-1)+\alpha}$	$\frac{\alpha}{(N-1)+\alpha}$	$\frac{(N-1)+\alpha^2}{((N-1)+\alpha)^2}$
$\alpha^{\frac{i-1}{N-1}}$	$\frac{\alpha^{\frac{1}{N-1}}-1}{\alpha^{\frac{N}{N-1}}-1}$	$\frac{\alpha^{\frac{N}{N-1}}-\alpha}{\alpha^{\frac{N}{N-1}}-1}$	$\frac{\left(\alpha^{\frac{2N}{N-1}}-1\right)/\left(\alpha^{\frac{2}{N-1}}-1\right)}{\left(\alpha^{\frac{N}{N-1}}-1\right)^2/\left(\alpha^{\frac{1}{N-1}}-1\right)^2}$
Cigar	$\frac{1}{(N-1)\alpha+1}$	$\frac{\alpha}{(N-1)\alpha+1}$	$\frac{(N-1)\alpha^2+1}{((N-1)\alpha+1)^2}$
$\alpha \cdot i$	$\frac{1}{N(N+1)/2}$	$\frac{1}{(N+1)/2}$	$\frac{\frac{1}{6}N(N+1)(2N+1)}{\left(N(N+1)/2\right)^2}$

where $X_k = \boldsymbol{m}^{(t)} + \sigma^{(t)} Z_k$.

The following lemma provides the expression of the conditional expectation of the weight function.

Lemma 3.7. *Let $X \sim \mathcal{N}(\boldsymbol{m}, \sigma^2 \mathbf{I})$ and $(X_i)_{i=1}^{\lambda}$ be the i.i.d. copies of X. Let $c_f(t) = \Pr[f(X) < t]$ be the cumulative density function of the function value $f(X)$. Then, we have for any $i, j \in [\![1, \lambda]\!]$, $i \neq j$,*

$$\mathbb{E}_{X_k, k\neq i}[W(i; (X_k)_{k=1}^{\lambda}) \mid f(X_i)] = u_1(c_f(f(X_i)))$$
$$\mathbb{E}_{X_k, k\neq i}[W(i; (X_k)_{k=1}^{\lambda})^2 \mid f(X_i)] = u_2(c_f(f(X_i)))$$
$$\mathbb{E}_{X_k, k\neq i, k\neq j}[W(i; (X_k)_{k=1}^{\lambda}) W(j; (X_k)_{k=1}^{\lambda}) \mid f(X_i), f(X_j)] = u_3(c_f(f(X_i)), c_f(f(X_j))) \enspace ,$$

where

$$u_1(p) = \sum_{k=1}^{\lambda} w_k \frac{(\lambda-1)!}{(k-1)!(\lambda-k)!} p^{k-1}(1-p)^{\lambda-k} \enspace , \tag{30}$$

$$u_2(p) = \sum_{k=1}^{\lambda} w_k^2 \frac{(\lambda-1)!}{(k-1)!(\lambda-k)!} p^{k-1}(1-p)^{\lambda-k} \enspace . \tag{31}$$

$$\begin{aligned} u_3(p,q) = \sum_{k=1}^{\lambda-1} \sum_{l=k+1}^{\lambda} w_k w_l \frac{(\lambda-2)!}{(k-1)!(l-k-1)!(\lambda-l)!} \\ \times \min(p,q)^{k-1} |q-p|^{l-k-1} (1-\max(p,q))^{\lambda-l} \enspace . \end{aligned} \tag{32}$$

Thanks to Lemma 3.7 and the fact that $(X_k)_{k=1}^{\lambda}$ are i.i.d., we can further rewrite the normalized quality gain by taking the iterated expectation $\mathbb{E} = \mathbb{E}_{X_i}\mathbb{E}_{X_k, k\neq i}$ as

$$\begin{aligned} &\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) \\ &= -\bar{\sigma}\lambda \mathbb{E}[u_1(c_f(f(X)))\mathbf{e}^{\mathrm{T}} Z] \\ &\quad - \frac{\bar{\sigma}^2 \lambda}{2} \mathbb{E}\left[u_2(c_f(f(X))) \frac{Z^{\mathrm{T}} \mathbf{A} Z}{\mathrm{Tr}(\mathbf{A})}\right] \\ &\quad - \frac{\bar{\sigma}^2 (\lambda-1)\lambda}{2} \mathbb{E}\left[u_3(c_f(f(X)), c_f(f(\tilde{X}))) \frac{Z^{\mathrm{T}} \mathbf{A} \tilde{Z}}{\mathrm{Tr}(\mathbf{A})}\right] \\ &= -\bar{\sigma}\lambda \mathbb{E}[u_1(c_f(f(X)))\mathbf{e}^{\mathrm{T}} Z] - \frac{\bar{\sigma}^2 \lambda}{2} \mathbb{E}[u_2(c_f(f(X)))] \\ &\quad - \frac{\bar{\sigma}^2 \lambda}{2} \mathbb{E}\left[u_2(c_f(f(X))) \left(\frac{Z^{\mathrm{T}} \mathbf{A} Z}{\mathrm{Tr}(\mathbf{A})} - 1\right)\right] \\ &\quad - \frac{\bar{\sigma}^2 (\lambda-1)\lambda}{2} \mathbb{E}\left[u_3(c_f(f(X)), c_f(f(\tilde{X}))) \frac{Z^{\mathrm{T}} \mathbf{A} \tilde{Z}}{\mathrm{Tr}(\mathbf{A})}\right] \enspace . \end{aligned} \tag{33}$$

Here Z and $\tilde{Z}$ are independent and N-multivariate normally distributed with zero mean and identity covariance matrix and $X = \boldsymbol{m} + \sigma Z$ and $\tilde{X} = \boldsymbol{m} + \sigma \tilde{Z}$, where $\sigma = \bar{\sigma}\|\nabla f(\boldsymbol{m})\|/(c_m \mathrm{Tr}(\mathbf{A}))$.

The following lemma shows that the second term on the right-most side of (33) depends only on the weights $(w_k)_{k=1}^{\lambda}$.

Lemma 3.8. $\lambda \mathbb{E}[u_2(c_f(f(X)))] = \sum_{i=1}^{\lambda} w_i^2$.

The following two lemmas show that the third and the fourth terms on the right-most size of (33) fade away as $\mathrm{Tr}(\mathbf{A}_N^2)/\mathrm{Tr}(\mathbf{A}_N)^2 \to 0$.

Lemma 3.9.

$$\sup_{\boldsymbol{m}\in\mathbb{R}^N} \left| \mathbb{E}\left[u_2(c_f(f(X))) \left(\frac{Z^{\mathrm{T}} \mathbf{A} Z}{\mathrm{Tr}(\mathbf{A})} - 1\right)\right]\right| \leqslant 2^{\frac{1}{2}} \left(\frac{1}{\lambda}\sum_{k=1}^{\lambda} w_k^4\right)^{\frac{1}{2}} \left(\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right)^{\frac{1}{2}} \enspace .$$

Lemma 3.10.

$$\sup_{\boldsymbol{m}\in\mathbb{R}^N} \left| \mathbb{E}\left[u_3(c_f(f(X)), c_f(f(\tilde{X}))) \frac{Z^{\mathrm{T}} \mathbf{A} \tilde{Z}}{\mathrm{Tr}(\mathbf{A})}\right]\right| \leqslant \left(\frac{2}{\lambda(\lambda-1)} \sum_{k=1}^{\lambda}\sum_{l=k+1}^{\lambda} w_k^2 w_l^2\right)^{\frac{1}{2}} \left(\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right)^{\frac{1}{2}} \enspace .$$

Finally we deal with the first term on the right-most side of (33). For this purpose, we introduce an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{e} = (1, 0, \ldots, 0)$, and let $\mathcal{N} = \mathbf{Q}Z$. Since Z is N-multivariate standard normally distributed, so is its orthogonal transformation $\mathcal{N}$. Hereafter, for any $x \in \mathbb{R}^N$ the ith coordinate of x is denoted by $[x]_i$.

Lemma 3.11.

$$\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |\mathbb{E}[u_1(c_f(f(X)))[\mathcal{N}]_1] - \mathbb{E}[u_1(\Phi([\mathcal{N}]_1))[\mathcal{N}]_1]| \leqslant 3(\lambda-1) \max_{k\in[\![1,\lambda-1]\!]} |w_{k+1} - w_k| \left[\frac{1}{4\pi}\left(\frac{\bar{\sigma}}{c_m}\right)^2 \frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right]^{\frac{1}{3}} \enspace .$$

Lemma 3.12. *Let $\mathcal{N}_{1:\lambda} \leqslant \ldots \leqslant \mathcal{N}_{\lambda:\lambda}$ be the normal order statistics. Then, $\lambda \mathbb{E}[u_1(\Phi([\mathcal{N}]_1))[\mathcal{N}]_1] = \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]$.*

We finalize the proof of Theorem 3.4. From (33), we have

$$\begin{aligned} &\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \left|\bar{\phi}(\boldsymbol{m}, \bar{\sigma}) - \left(-\bar{\sigma}\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] - \frac{\bar{\sigma}^2}{2}\sum_{i=1}^{\lambda} w_i^2\right)\right| \\ &\leqslant \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \left|-\bar{\sigma}\lambda\mathbb{E}[u_1(c_f(f(X)))\mathbf{e}^{\mathrm{T}} Z] - \left(-\bar{\sigma}\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]\right)\right| \\ &+ \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \left|-\frac{\bar{\sigma}^2\lambda}{2}\mathbb{E}[u_2(c_f(f(X)))] - \left(-\frac{\bar{\sigma}^2}{2}\sum_{i=1}^{\lambda} w_i^2\right)\right| \\ &+ \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \left|-\frac{\bar{\sigma}^2\lambda}{2}\mathbb{E}\left[u_2(c_f(f(X)))\left(\frac{Z^{\mathrm{T}}\mathbf{A}Z}{\mathrm{Tr}(\mathbf{A})} - 1\right)\right]\right| \\ &+ \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \left|-\frac{\bar{\sigma}^2(\lambda-1)\lambda}{2}\mathbb{E}\left[u_3(c_f(f(X)), c_f(f(\tilde{X})))\frac{Z^{\mathrm{T}}\mathbf{A}\tilde{Z}}{\mathrm{Tr}(\mathbf{A})}\right]\right| \enspace . \end{aligned}$$

In light of Lemmas 3.11 and 3.12, the first term is upper bounded by

$$\frac{3}{(4\pi)^{\frac{1}{3}}} \frac{\bar{\sigma}^{\frac{5}{3}}}{c_m^{\frac{2}{3}}} \lambda(\lambda-1) \max_{k\in[\![1,\lambda-1]\!]} |w_{k+1} - w_k| \left(\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right)^{\frac{1}{3}} \enspace . \tag{34}$$

Thanks to lemma 3.8, the second term is zero. The third term and the fourth term are bounded using Lemmas 3.9 and 3.10, respectively, as

$$\frac{1}{2^{\frac{1}{2}}}\bar{\sigma}^2\left(\lambda\sum_{k=1}^{\lambda}w_k^4\right)^{\frac{1}{2}}\left(\frac{\operatorname{Tr}(\mathbf{A}^2)}{\operatorname{Tr}(\mathbf{A})^2}\right)^{\frac{1}{2}} \quad \text{and}$$

$$\frac{1}{2^{\frac{1}{2}}}\bar{\sigma}^2\left((\lambda-1)\lambda\sum_{k=1}^{\lambda}\sum_{l=k+1}^{\lambda}w_k^2w_l^2\right)^{\frac{1}{2}}\left(\frac{\operatorname{Tr}(\mathbf{A}^2)}{\operatorname{Tr}(\mathbf{A})^2}\right)^{\frac{1}{2}} .$$

This completes the proof.

4. SIMULATION

We conduct experiments to see (i) how the empirical normalized quality gain deviates from the asymptotic normalized quality gain (9) on a finite dimensional quadratic function, (ii) how it depends on the Hessian $\mathbf{A}$, and (iii) how it changes when we change c_m while fixing $\bar{\sigma} \times c_m$.

To estimate the normalized quality gain, we run T iterations of the algorithm (1) with scale-invariant step-size (16) and average the empirical quality gain at each iteration, namely,

$$\frac{0.9}{T}\sum_{t=T/10}^{T-1}\bar{\phi}(\boldsymbol{m}^{(t)},\bar{\sigma}) \ . \tag{35}$$

We set $T = 10000$ and run 10 trials with the initial mean $\boldsymbol{m}^{(0)}$ taken from the uniform distribution on the unit sphere surface[7]. We plot the median ($\times$) and the 10%- and 90%-tile range (shaded area in same color, always thin and often not even visible) over the trials, along with the asymptotic normalized quality gain $\bar{\phi}_\infty$ ($\times$) and the theoretical bound with $c_m = 1$ (outside shaded area) and $c_m = \infty$ (inside shaded area). The results are summarized in Figure 4. Note that the normalized quality gain is divided by λ, i.e., per-sample normalized quality gain is displayed, and the normalized step-size is divided by the optimal normalized step-size $\bar{\sigma}^* = \bar{\sigma}^*((w_k)_{k=1}^{\lambda})$ given in (16) (with a slight abuse of notation) so that the peak of $\bar{\phi}_\infty$ is located at 1 and its value is around 0.5 for the optimal weights and 0.25 for the non-negative CMA-type weights.

Effect of A. We first focus on the results with $c_m = 1$ (the default setting). The empirical normalized quality gain gets closer to the normalized quality gain derived for the infinite dimensional quadratic function as N increases. The approach of the empirical normalized quality gain to the theory is the fastest for the sphere function ($\mathbf{A} = \mathbf{I}$). For convex quadratic functions with the same condition number of $\alpha = 10^6$, the speed of the convergence of the *normalized* quality gain to $\bar{\phi}_\infty$ as $N \to \infty$ is the fastest for the cigar function, and the slowest for the discus function. This reflects the upper bound derived in Theorem 3.4 that depends on the ratio $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2$, whose value is summarized in Table 1 for the functions tested here. For the cigar function the (normalization) ratio is close to $1/(N-1)$, while for the discus function the ratio is very close to 1 for $N \ll \alpha$ and we do not observe significant difference between results on different N.

[7]We checked the dependency of the estimated normalized quality gain on the number of iterations T. Comparing the median values over 10 trials obtained with $T = 1000$ and $T = 10000$, the maximum difference was 0.0004 for $\bar{\sigma}/\bar{\sigma}^* \leqslant 0.1$, and 0.009 for $\bar{\sigma}/\bar{\sigma}^* \leqslant 2$ among all settings (all functions, all dimensions, both types of weights, and all values of c_m).

Effect of c_m. On the sphere function, we observe that a larger c_m leads to a better empirical normalized quality gain if $\bar{\sigma}c_m$ is fixed. The reason is as follows. On the sphere function, the negative gradient everywhere points to the global optimum. The function is locally approximated by a linear function whose normal vector is given by the gradient of the function. When c_m is large and hence the step-size is small, the ranking of the candidate solutions tends to admit the ranking of the solutions on the linear function. Then, it tends to provide a good estimate of the gradient. Therefore, a smaller step size, i.e., a larger c_m, leads to better progress. On the other quadratic functions, we observe a similar tendency when $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2$ is small enough. However, if λ is relatively large or $\operatorname{Tr}(\mathbf{A}^2)/\operatorname{Tr}(\mathbf{A})^2$ is relatively large, particularly on the discus function, a large c_m tends to result in a poor convergence performance. On a quadratic function in general, its gradient does not point to the optimum. If σ is sufficiently small but σc_m is relatively large, the mean vector is updated in the negative gradient direction and overshoots since the actual step-size σc_m is too large. Consequently, σ will become even smaller in practice. This behavior will be more emphasized if the condition number of the Hessian is higher.

Effect of w_k. We clearly see the difference between the normalized quality gain response with the optimal weights and the non-negative weights used in the CMA-ES. Compared with the curves for the optimal weights, the curves for the CMA-type weights tend to be closer to the asymptotic normalized quality gain. The difference is emphasized when the learning rate c_m for the mean vector update is smaller than one. Differently from the tendency we observe for the optimal weights, a smaller c_m results in a higher empirical quality gain.

The optimal weights are symmetric around zero, implying that they assume the symmetry of the sorted candidate solutions. Such a symmetry does not hold for a general convex quadratic function in a finite dimensional search space unless the step-size is sufficiently small. Since the symmetry can not be assumed in practice, using the negative part of the weights leads to unpromising behavior when c_m is small, i.e., $\bar{\sigma}$ is large.

5. FURTHER DISCUSSION

Tightness of the Normalized Quality Gain Bound. As we see in Figures 4, the theoretically derived bound (Theorem 3.4) is not tight, in particular for the interesting area where the normalized quality gain has a peak.

The bound (20) has two terms. The first term is mainly from the approximation error of the cumulative density function of $f(X)$ (scaled by the right factor) for $X \sim \mathcal{N}(\boldsymbol{m}, \sigma\mathbf{I})$ by the cumulative density function of the univariate standard normal distribution $\mathcal{N}$. In the proof of Lemma 3.11, we use the Chebyshev's inequality to bound the error. However, since we know the distribution of $\mathcal{N}$ and $f(X)$, we may be able to improve the first term by exploiting the distributions. Since the first term dominates if c_m is fixed, a better bound for the first term will lead to a tighter bound.

The second term of (20) comes from the quadratic terms of (33). It bounds the deviation of the normalized quality gain due to a random walk of $\boldsymbol{m}$ on the equal function

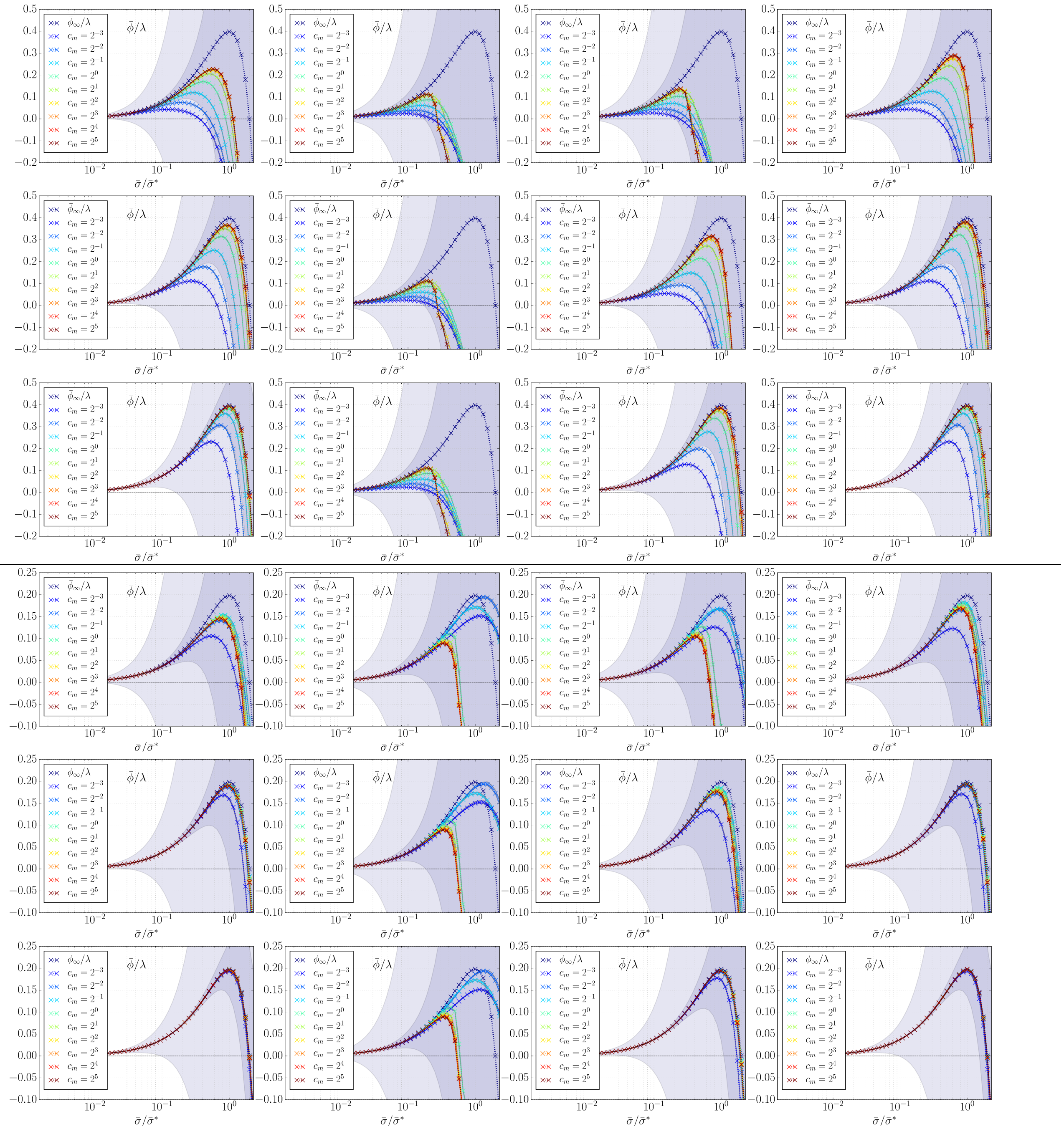


Figure 4: Empirical normalized quality gain on four convex quadratic functions, Sphere, Discus, Ellipsoid and Cigar (from left to right) of dimension $N = 10$, 100 and 1000 (from top to bottom). The optimal weights (top) and the (nonnegative) CMA-type weights are used and $\lambda = 10$.

value surface (hyperellipsoid). It becomes almost independent of selection (ranking) of candidate solutions as N goes sufficiently large. In the proofs of Lemmas 3.9 and 3.10 we simply use the Cauchy-Schwarz inequality to evaluate the effect of selection (weight functions u_2 and u_3) and the noise due to stochastic sampling of candidate solutions separately. The second term gives the best upper bound that is achieved when $c_m \to \infty$, though it is not a realistic setting. Even if $c_m \to \infty$ is considered, we can not yet obtain the $O(\lambda/N)$ scale up of the quality gain for $\lambda \in o(N)$, but we obtain $\lambda \in o(N^{\frac{1}{2}})$ for $(\mu/\mu_I, \lambda)$-ES with fixed μ/λ ratio. To answer the question if we can achieve $O(\lambda/N)$ scale up for $\lambda \in o(N)$, we need to improve both the first and second terms of the bound (20).

Non-Isotropic Gaussian Sampling. The isotropic covariance matrix of the multivariate normal sampling distribution is considered above. We can generalize all of the results in this paper to an arbitrary positive definite symmetric covariance matrix $\mathbf{C}$ by considering the affine transformation of the search space. Let $f : x \mapsto (x - x^*)^{\mathrm{T}} \mathbf{A}(x - x^*)$, and consider the coordinate transformation $x \mapsto y = \mathbf{C}^{-\frac{1}{2}} x$. In the latter coordinate system the function f can be written as $f(x) = \bar{f}(y) = (y - \mathbf{C}^{-\frac{1}{2}} x^*)^{\mathrm{T}} (\mathbf{C}^{\frac{1}{2}} \mathbf{A} \mathbf{C}^{\frac{1}{2}})(y - \mathbf{C}^{-\frac{1}{2}} x^*)$. The multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \sigma^2 \mathbf{C})$ is transformed into $\mathcal{N}(\mathbf{C}^{-\frac{1}{2}} \boldsymbol{m}, \sigma^2 \mathbf{I})$ by the same transformation. Then, it is easy to prove that the quality gain on the function f given the parameter $(\boldsymbol{m}, \sigma, \mathbf{C})$ is equivalent to the quality gain on the function $\bar{f}$ given $(\mathbf{C}^{-\frac{1}{2}} \boldsymbol{m}, \sigma^2, \mathbf{I})$. The normalization factor $g(\boldsymbol{m})$ of the quality gain and the normalized step-size are then rewritten as

$$g(\boldsymbol{m}) = \frac{\|\mathbf{C}^{\frac{1}{2}} \mathbf{A}(\boldsymbol{m} - x^*)\|^2}{f(\boldsymbol{m}) \operatorname{Tr}(\mathbf{C}^{\frac{1}{2}} \mathbf{A} \mathbf{C}^{\frac{1}{2}})} \ , \tag{36}$$

$$\bar{\sigma} = \frac{\sigma c_m \operatorname{Tr}(\mathbf{C}^{\frac{1}{2}} \mathbf{A} \mathbf{C}^{\frac{1}{2}})}{\|\mathbf{C}^{\frac{1}{2}} \mathbf{A}(\boldsymbol{m} - x^*)\|} \ . \tag{37}$$

Interpretation of Dynamics. The asymptotic quality gain depends on $\boldsymbol{m}$ through $g(\boldsymbol{m})$. In practice, we observe near worst case performance with $g(\boldsymbol{m}) \approx 2 d_N(\mathbf{A}) / \operatorname{Tr}(\mathbf{A})$, which implies that $\boldsymbol{m} - x^*$ is almost parallel to the eigenspace corresponding to the smallest eigenvalue $d_N(\mathbf{A})$ of the Hessian matrix. We provide an intuition to explain this behavior, which will be useful to understand the algorithm, even though the argument is not fully rigorous.

We consider the weighted recombination ES (6) with scale-invariant step-size (16). Lemma 3.11 implies that the order of the function values $f(X_i)$ coincide with the order of $[\mathcal{N}_i]_1 = \mathbf{e}^{\mathrm{T}} (X_i - \boldsymbol{m}^{(t)}) / \sigma^{(t)}$, where $\mathbf{e} = \nabla f(\boldsymbol{m}^{(t)}) / \|\nabla f(\boldsymbol{m}^{(t)})\|$. This is because if $d \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $d^{\mathrm{T}} \mathbf{A} d / \operatorname{Tr}(\mathbf{A})$ in (28) almost surely converges to one by the strong law of large numbers. It means that the function value of the candidate solutions is determined solely by the first component on the right-hand side of (28), that is, $\mathbf{e}^{\mathrm{T}} (X_i - \boldsymbol{m}^{(t)}) / \sigma^{(t)}$.

Since the ranking of the function only depends on $\mathbf{e}^{\mathrm{T}} (X_i - \boldsymbol{m}^{(t)}) / \sigma^{(t)}$, one can rewrite the update of the mean vector as

$$\begin{aligned} \boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sigma^{(t)} \textstyle\sum_{i=1}^{\lambda} w_i \mathcal{N}_{i:\lambda}(0,1) \cdot \mathbf{e} \\ + c_m \sigma^{(t)} \mu_w^{-\frac{1}{2}} \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}) \ , \end{aligned} \tag{38}$$

where $\mu_w = \left(\sum_{i=1}^{\lambda} w_i^2 \right)^{-1}$, $\mathcal{N}_{i:\lambda}(0,1)$ are the ith order statistics from λ population of $\mathcal{N}(0,1)$, and $\mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}})$ is the normal distributed random vector with mean vector $\mathbf{0}$ and the degenerated covariance matrix $\mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}$. It implies that the mean vector moves along the gradient direction with the distribution $c_m \sigma^{(t)} \sum_{i=1}^{\lambda} w_i \mathcal{N}_{i:\lambda}(0,1)$, while it moves randomly in the subspace orthogonal to the gradient with the distribution $c_m \sigma^{(t)} \mu_w^{-\frac{1}{2}} \mathcal{N}(\mathbf{0}, \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}})$.

If the function is the spherical function, $\mathbf{A} \propto \mathbf{I}$, the mean vector does a symmetric, unbiased random walk on the surface of a hypersphere while the radius of the hypersphere gradually decreases due to the second term on (38). If the function is a general convex quadratic function, $\mathbf{A} \not\propto \mathbf{I}$, the corresponding random walk on the surface of a hyperellipsoid becomes biased. Then, $\boldsymbol{m} - x^*$ tends to be parallel to the eigenspace corresponding to the smallest eigenvalue $d_N(\mathbf{A})$, which means that the quality gain is close to the worst case of $d_N(\mathbf{A}) / \operatorname{Tr}(\mathbf{A})$[8].

Geometric Interpretation of the Optimal Situation. On the sphere function, we know that the optimal step-size puts the algorithm in the situation where $f(\boldsymbol{m})$ improves twice as much by $\boldsymbol{m}$ moving towards the optimum as it deteriorates by $\boldsymbol{m}$ moving randomly in the subspace orthogonal to the gradient direction [24]. On a general convex quadratic function, we find the analogous result. From (33) and lemmas in Section 3.4, the first term of the asymptotic normalized quality gain (10), i.e. $-\bar{\sigma} \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}]$, is due to the movement of $\boldsymbol{m}$ towards the negative gradient direction, and the second term, i.e. $-\frac{\bar{\sigma}^2}{2} \sum_{i=1}^{\lambda} w_i^2$, is due to the random walk in the orthogonal subspaces[9]. The asymptotic normalized quality gain is maximized when the normalized step-size is set such that $-\bar{\sigma} \sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}] = \bar{\sigma}^2 \sum_{i=1}^{\lambda} w_i^2$. That is, the amount of the decrease of $f(\boldsymbol{m})$ by $\boldsymbol{m}$ moving into the negative gradient direction is twice greater than the increase of $f(\boldsymbol{m})$ by $\boldsymbol{m}$ moving in its orthogonal subspace.

Linear Convergence. We can derive the linear convergence of the algorithm where at each iteration, the step-size is proportional to $\|\nabla f(\boldsymbol{m})\| / c_m \operatorname{Tr}(\mathbf{A})$ from our quality gain analysis. Since the algorithm is stochastic, there are several ways to define the linear convergence. Here, we define the linear convergence of the algorithms by

$$\lim_{t \to \infty} -\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\ln \frac{f(\boldsymbol{m}^{(i+1)})}{f(\boldsymbol{m}^{(i)})} \right] = \mathrm{CR} > 0 \ . \tag{39}$$

If we consider the weighted recombination evolution strategy with the step-size (16), using the normalized log quality gain, we can rewrite the above definition as

$$-\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\ln \frac{f(\boldsymbol{m}^{(i+1)})}{f(\boldsymbol{m}^{(i)})} \right] = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[g(\boldsymbol{m}^{(i)}) \bar{\phi}_{\ln}(\boldsymbol{m}^{(i)}, \bar{\sigma}) \right] .$$

The convergence CR is lower bounded by using Theorem 3.4. Using the inequality $g(\boldsymbol{m}) \geqslant 2 d_N(\mathbf{A}) / \operatorname{Tr}(\mathbf{A})$, we can lower bound each term on the right-hand side of the above equality as

$$g(\boldsymbol{m}^{(t)}) \bar{\phi}_{\ln}(\boldsymbol{m}^{(t)}, \bar{\sigma}) \geqslant \frac{2 d_N(\mathbf{A})}{\operatorname{Tr}(\mathbf{A})} \bar{\phi}_{\ln}(\boldsymbol{m}^{(t)}, \bar{\sigma}) \ .$$

Then, note that $\bar{\phi}_{\ln}(\boldsymbol{m}, \bar{\sigma}) \geqslant \bar{\phi}(\boldsymbol{m}, \bar{\sigma})$. Letting the right-hand side of the inequality (20) be denoted by $\epsilon(\mathbf{A}, \bar{\sigma}, (w_k)_{k=1}^{\lambda})$, we have

$$\bar{\phi}_{\ln}(\boldsymbol{m}^{(t)}, \bar{\sigma}) \geqslant \bar{\phi}(\boldsymbol{m}, \bar{\sigma}) \geqslant \bar{\phi}_{\infty}(\bar{\sigma}, (w_k)_{k=1}^{\lambda}) - \epsilon(\mathbf{A}, \bar{\sigma}, (w_k)_{k=1}^{\lambda}) \ .$$

[8] The reason may be explained as follows. It is illustrated in Figure 3 that the optimal step-size is the largest when $\boldsymbol{m}$ is on the shortest axis of the hyperellipsoid, and the smallest when $\boldsymbol{m}$ is on the longest axis. It implies that the progress in one step is the largest in the short axis direction (parallel to the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$), and the smallest in the long axis direction (parallel the eigenvector corresponding to the largest eigenvalue of $\mathbf{A}$). It implies the situation quickly becomes closer to the worst case, while it takes longer iterations to escape from near worst situation. Therefore, we observe near worst situation in practice.

[9] More precisely, the second term comes from the quadratic term in (28) that contains the information in the gradient direction as well. However, the above statement is true in the limit $N \to \infty$.

Altogether, we obtain

$$\mathrm{CR} \geqslant \frac{2d_N(\mathbf{A})}{\mathrm{Tr}(\mathbf{A})} \left(\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda) - \epsilon(\mathbf{A}, \bar{\sigma}, (w_k)_{k=1}^\lambda) \right) \ .$$

Therefore, if $\bar{\phi}_\infty(\bar{\sigma}, (w_k)_{k=1}^\lambda) > \epsilon(\mathbf{A}, \bar{\sigma}, (w_k)_{k=1}^\lambda)$, the convergence rate CR > 0 and the algorithm converges. For choices of λ and N that were discussed previously in the paper, we know that $\epsilon(\mathbf{A}, \bar{\sigma}, (w_k)_{k=1}^\lambda)$ goes to zero such that the condition will be satisfied.

Related Work. Reference [19] studied the $(1+1)$-ES with the one-fifth success rule on a convex quadratic function with Hessian

$$\mathbf{A} = \frac{1}{2} \operatorname{diag}(\underbrace{\alpha, \dots, \alpha}_{\lfloor N\theta \rfloor}, \underbrace{1, \dots, 1}_{N - \lfloor N\theta \rfloor}) \ , \tag{40}$$

where $\alpha \geqslant 1$ is the condition number, $\theta \in [0, 1]$ determines the proportion of the number of short axes of the ellipsoid. To analyze the algorithmic behavior, the so-called *localization parameter*, ζ, that is the distance to the optimum in the last $N - \lfloor N\theta \rfloor$ dimensional subspace divided by the distance to the optimum in the first $\lfloor N\theta \rfloor$ times α. In the reference, it is proved that for $\alpha \in poly(N)$ and $\theta = 1/2$, if the initialization satisfies $\sigma^{(0)} = \Theta(f(\boldsymbol{m})^{\frac{1}{2}}/N/\alpha)$ and $\zeta^{-1} = O(1)$, then the number of iterations to reduce the initial f-value to a 2^{-b}-fraction for $b = poly(N)$ is $\Theta(b\alpha N)$. It translates as that the convergence rate is $\Theta(1/(\alpha N))$. It almost matches the worst case of the quality gain, $\Omega(d_N(\mathbf{A})/\operatorname{Tr}(\mathbf{A})) = \Omega(2/((\alpha+1)N))$. The convergence rate of $O(1/N)$ is also proved on a general quadratic function with a bounded condition number, $\alpha \in O(1)$. However, the influence of the condition number α to the convergence rate is missing.

References [5] and [16] studied ES with intermediate recombination and ES with weighted recombination, respectively, on the same type of quadratic functions with Hessian (40). Their results, progress rate and quality gain, depend on the localization parameter, the steady-state value of which is then analyzed to obtain the steady-state quality gain. References [12, 13] studied the progress rate and the quality gain on the general convex quadratic model as we already mentioned in the paper. Our study is differentiated from these results by a mathematical rigorousness. On the other hands, in these references, the dynamics of the step-size control mechanisms, self-adaptation and cumulative step-size adaptation, is analyzed. These are beyond the scope of this paper and we do not discuss further on this topic.

6. ACKNOWLEDGEMENTS

This work is partially supported by JSPS KAKENHI Grant Number 15K16063.

7. REFERENCES

[1] Y. Akimoto, A. Auger, and N. Hansen. Convergence of the continuous time trajectories of isotropic evolution strategies on monotonic $\mathcal{C}^2$-composite functions. In *Problem Solving from Nature - PPSN XII*, pages 42–51, 2012.

[2] Y. Akimoto and N. Hansen. Online model selection for restricted covariance matrix adaptation. In *Parallel Problem Solving from Nature - PPSN XIV*, pages 3–13, 2016.

[3] Y. Akimoto and N. Hansen. Projection-based restricted covariance matrix adaptation for high dimension. In *Genetic and Evolutionary Computation Conference, GECCO*, pages 197–204, 2016.

[4] D. V. Arnold. Optimal weighted recombination. In *Foundations of Genetic Algorithms*, pages 215–237, 2005.

[5] D. V. Arnold. On the use of evolution strategies for optimising certain positive definite quadratic forms. In *Genetic and Evolutionary Computation Conference, GECCO*, pages 634–641, 2007.

[6] A. Auger. *Analysis of Comparison-based Stochastic Continuous Black-Box Optimization Algorithms.* Habilitation, Universitè Paris-Sud, 2015.

[7] A. Auger, D. Brockhoff, and N. Hansen. Mirrored sampling in evolution strategies with weighted recombination. In *Genetic and Evolutionary Computation Conference, GECCO*, pages 861–868, 2011.

[8] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In *Genetic and Evolutionary Computation Conference, GECCO*, pages 445–452, 2006.

[9] A. Auger and N. Hansen. Linear convergence of comparison-based step-size adaptive randomized search via stability of markov chains. *SIAM Journal on Optimization*, 26(3):1589–1624, 2016.

[10] H.-G. Beyer. Towards a theory of 'evolution strategies': Results for (1+, λ)-strategies on (nearly) arbitrary fitness functions. In *Parallel Problem Solving from Nature - PPSN III*, pages 58–67, 1994.

[11] H.-G. Beyer. *The Theory of Evolution Strategies.* Natural Computing Series. Springer-Verlag, 2001.

[12] H.-G. Beyer and M. Hellwig. The dynamics of cumulative step size adaptation on the ellipsoid model. *Evolutionary Computation*, 24(1):25–57, 2016.

[13] H.-G. Beyer and A. Melkozerov. The dynamics of self-adaptive multirecombinant evolution strategies on the general ellipsoid model. *IEEE Transactions on Evolutionary Computation*, 18(5):764–778, 2014.

[14] A. DasGupta. *Asymptotic theory of statistics and probability.* Springer Science & Business Media, 2008.

[15] L. Devroye. *Non-Uniform Random Variate Generation.* Springer New York, 1986.

[16] S. Finck and H.-G. Beyer. Weighted recombination evolution strategy on a class of pdqf's. In *Foundations of Genetic Algorithms, FOGA*, pages 1–12, 2009.

[17] N. Hansen and A. Auger. Principled design of continuous stochastic search: From theory to practice. In Y. Borenstein and A. Moraglio, editors, *Theory and Principled Methods for the Design of Metaheuristics.* Springer, 2014.

[18] N. Hansen and S. Kern. Evaluating the cma evolution strategy on multimodal test functions. In *Parallel Problem Solving from Nature - PPSN VIII*, pages 282–291, 2004.

[19] J. Jägersküpper. How the $(1+1)$ es using isotropic mutations minimizes positive definite quadratic forms. *Theoretical Computer Science*, 361(1):38–56, 2006.

[20] M. Jebalia and A. Auger. Log-linear convergence of the scale-invariant $(\mu/\mu_w, \lambda)$-es and optimal μ for

intermediate recombination for large population sizes. In *Parallel Problem Solving from Nature - PPSN XI*, pages 52–62, 2010.

[21] M. Jebalia and A. Auger. Log-linear convergence of the scale-invariant $(\mu/\mu_w, \lambda)$-ES and optimal mu for intermediate recombination for large population sizes. Research Report RR-7275, Jun 2010.

[22] M. Jebalia, A. Auger, and P. Liardet. Log-linear convergence and optimal bounds for the (1+1)-es. In *Evolution Artificielle (EA '07)*, pages 207–218, 2008.

[23] I. Loshchilov. A computationally efficient limited memory cma-es for large scale optimization. In *Genetic and Evolutionary Computation Conference, GECCO*, pages 397–404, 2014.

[24] I. Rechenberg. *Evolutionsstrategie '94*. Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1994.

[25] R. Ros and N. Hansen. A simple modification in cma-es achieving linear time and space complexity. In *Parallel Problem Solving from Nature - PPSN X*, pages 296–305, 2008.

[26] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In *Parallel Problem Solving from Nature - PPSN IX*, pages 21–31, 2006.

APPENDIX

A. PROOFS

In the following, let $H_N = \frac{\bar\sigma}{c_m}[\mathcal{N}]_1 + \frac{\bar\sigma^2}{2c_m^2}\frac{\mathcal{N}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\mathcal{N}}{\mathrm{Tr}(\mathbf{A})}$. Let $\tilde{\mathcal{N}}$, $\tilde Z$, $\tilde X$, $\tilde H_N$ be the i.i.d. copies of $\mathcal{N}$, Z, X, H_N, respectively. Moreover, let c_N be the cumulative density function of H_N.

Proof of Proposition 3.1. Let $\boldsymbol{d} = c_m \sum_{i=1}^{\lambda} W(i; (\boldsymbol{m}+\sigma Z_k)_{k=1}^{\lambda}) Z_i$, where $(Z_i)_{i=1}^{\lambda}$ are independent and N-variate standard normally distributed random vectors. Then,

$$
\begin{aligned}
&\phi(x^* + (\boldsymbol{m} - x^*), \sigma)\\
&= \phi(\boldsymbol{m}, \sigma)\\
&= -\frac{\mathbb{E}[g(\boldsymbol{m} + \sigma \boldsymbol{d})] - g(\boldsymbol{m})}{g(\boldsymbol{m})}\\
&= -\frac{\mathbb{E}[f(\boldsymbol{m} + \sigma \boldsymbol{d} - x^*)] - f(\boldsymbol{m} - x^*)}{f(\boldsymbol{m} - x^*)}\\
&= -\frac{\alpha^{-n}\mathbb{E}[f(\alpha\cdot(\boldsymbol{m} + \sigma \boldsymbol{d} - x^*))] - \alpha^{-n} f(\alpha\cdot(\boldsymbol{m} - x^*))}{\alpha^{-n} f(\alpha\cdot(\boldsymbol{m} - x^*))}\\
&= -\frac{\mathbb{E}[f(\alpha\cdot(\boldsymbol{m} + \sigma \boldsymbol{d} - x^*))] - f(\alpha\cdot(\boldsymbol{m} - x^*))}{f(\alpha\cdot(\boldsymbol{m} - x^*))}\\
&= -\frac{\mathbb{E}[g(x^* + \alpha\cdot(\boldsymbol{m} - x^*) + \alpha\sigma\boldsymbol{d})] - g(x^* + \alpha\cdot(\boldsymbol{m} - x^*))}{g(x^* + \alpha\cdot(\boldsymbol{m} - x^*))}\\
&= \phi(x^* + \alpha(\boldsymbol{m} - x^*), \alpha\sigma)\ .
\end{aligned}
$$

That is, the quality gain is scale invariant around $(x^*, 0)$. Moreover, the above equality implies that $\mathrm{argmax}_\sigma\, \phi(x^* + (\boldsymbol{m} - x^*), \sigma) = \mathrm{argmax}_\sigma\, \phi(x^* + \alpha(\boldsymbol{m} - x^*), \alpha\sigma)$, i.e., the optimal step-size at $x^* + \alpha(\boldsymbol{m} - x^*)$ is α times greater than the optimal step-size at $x^* + (\boldsymbol{m} - x^*)$. Therefore, the optimal step-size as a function of $\boldsymbol{m} - x^*$ is homogeneous with the exponent 1, i.e., $\sigma^*(\alpha\cdot(\boldsymbol{m} - x^*)) = \alpha\sigma^*(\boldsymbol{m} - x^*)$. □

Proof of Corollary 3.6. Let $\bar\sigma = \beta\bar\sigma_w^*$ for some $\beta \in [\epsilon, 2-\epsilon]$. Two terms depending on $\bar\sigma$ on the right-hand side of (20) divided by $\bar\phi_\infty(\bar\sigma, (w_k)_{k=1}^{\lambda})$ read

$$
\begin{aligned}
\frac{\bar\sigma^{\frac{5}{3}}}{\bar\phi_\infty(\beta\bar\sigma_w^*, (w_k)_{k=1}^{\lambda})} &= \frac{\beta^{\frac{2}{3}}/(1-\beta/2)}{(\sum_{i=1}^{\lambda} w_i^2)(\bar\sigma_w^*)^{\frac{1}{3}}}\\
&= \frac{\beta^{\frac{2}{3}}/(1-\beta/2)}{(\sum_{i=1}^{\lambda} w_i^2)^{\frac{2}{3}}(-\sum_{i=1}^{\lambda} w_i \mathbb{E}[\mathcal{N}_{i:\lambda}])^{\frac{1}{3}}}
\end{aligned}
\tag{41}
$$

and

$$
\frac{\bar\sigma^2}{\bar\phi_\infty(\bar\sigma, (w_k)_{k=1}^{\lambda})} = \frac{\beta/(1-\beta/2)}{(\sum_{i=1}^{\lambda} w_i^2)}\ . \tag{42}
$$

The three terms depending on λ and $(w_k)_{k=1}^{\lambda}$ on the right-hand side of (20) are respectively bounded as

$$
(\lambda^2 - \lambda)\max_{k\in[\![1,\lambda-1]\!]}|w_{k+1} - w_k| \leqslant \lambda^2 \max_{k\in[\![1,\lambda-1]\!]}|w_{k+1} - w_k| \tag{43}
$$

and

$$
\left(\lambda\sum_{k=1}^{\lambda} w_k^4\right)^{\frac{1}{2}} \leqslant \lambda^{\frac{1}{2}}\sum_{k=1}^{\lambda} w_k^2 \tag{44}
$$

$$
\left((\lambda-1)\lambda\sum_{k=1}^{\lambda}\sum_{l=k+1}^{\lambda} w_k^2 w_l^2\right)^{\frac{1}{2}} \leqslant \lambda\sum_{k=1}^{\lambda} w_k^2\ . \tag{45}
$$

Using the above equalities and inequalities, we obtain

$$
\begin{aligned}
&\frac{\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|\bar\phi(\boldsymbol{m},\bar\sigma) - \bar\phi_\infty(\bar\sigma,(w_k)_{k=1}^{\lambda})\right|}{\bar\phi_\infty(\bar\sigma,(w_k)_{k=1}^{\lambda})}\\
&\leqslant \frac{3\beta^{\frac{2}{3}}}{(4\pi)^{\frac{1}{3}} c_m^{\frac{2}{3}}(1-\frac{\beta}{2})}\left[\frac{\lambda^6 \max_{k\in[\![1,\lambda-1]\!]}|w_{k+1}-w_k|^3}{(\sum_{i=1}^{\lambda} w_i^2)^2(-\sum_{i=1}^{\lambda} w_i\mathbb{E}[\mathcal{N}_{i:\lambda}])}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right]^{\frac{1}{3}}\\
&+ \frac{\beta}{2^{\frac{1}{2}}(1-\frac{\beta}{2})}\left[(\lambda^{\frac{1}{2}}+\lambda)^2\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right]^{\frac{1}{2}}\ .
\end{aligned}
\tag{46}
$$

Under the condition (23), both the first and the second terms on the RHS of (46) converge to zero as $N \to \infty$. This completes the proof. □

Proof of Lemma 3.7. Since $(X_k)_{k=1}^{\lambda}$ are independently and normally distributed, the conditional probability of an event $\mathbb{I}\{f(X_k) < f(X_i)\} = 1$ for any $k \neq i$ given X_i is $c_f(f(X_i))$. Then, the probability of $\sum_{k=1}^{\lambda}\mathbb{I}\{f(X_k) \leqslant f(X_i)\}$ being a for $a \in [\![1,\lambda]\!]$ is given by the binomial probability mass

$$
\frac{(\lambda-1)!}{(a-1)!(\lambda-a)!}p^{a-1}(1-p)^{\lambda-a} \tag{47}
$$

with $p = c_f(f(X_i))$. Note that the sum of the above probabilities over $a \in [\![1,\lambda]\!]$ is one. It implies that $W(i;(X_k)_{k=1}^{\lambda})^\alpha$ takes the value of w_a with probability one. Then, for any $\alpha \geqslant 0$,

$$
\begin{aligned}
&\mathbb{E}_{X_k, k\neq i}[W(i;(X_k)_{k=1}^{\lambda})^\alpha \mid f(X_i)]\\
&\qquad = \sum_{k=1}^{\lambda} w_k^\alpha \frac{(\lambda-1)!}{(k-1)!(\lambda-k)!}p^{k-1}(1-p)^{\lambda-k}\ .
\end{aligned}
$$

Similarly, the joint probability of $\sum_{k=1}^{\lambda}\mathbb{I}\{f(X_k) \leqslant f(X_i)\}$ and $\sum_{k=1}^{\lambda}\mathbb{I}\{f(X_k) \leqslant f(X_j)\}$ is derived. Due to the symmetry of $W(i;(X_k)_{k=1}^{\lambda})W(j;(X_k)_{k=1}^{\lambda})$, we can assume w.l.g. that $f(X_i) \leqslant f(X_j)$. Then, the conditional joint probability of $\sum_{k=1}^{\lambda}\mathbb{I}\{f(X_k) \leqslant f(X_i)\} = a$ and $\sum_{k=1}^{\lambda}\mathbb{I}\{f(X_k) \leqslant$

$f(X_j)\} = b$ for $a, b \in [\![1, \lambda]\!]$ given X_i and X_j is expressed by the trinomial probability mass

$$\frac{(\lambda-2)!}{(a-1)!(b-a-1)!(\lambda-b)!} p^{a-1}(q-p)^{b-a-1}(1-q)^{\lambda-b} \quad (48)$$

with $p = c_f(f(X_i))$ and $q = c_f(f(X_j))$ if $a < b$, and zero otherwise. Note that the sum of the above probability over $1 \leqslant a < b \leqslant \lambda$ is 1. It implies that $W(i; (X_k)_{k=1}^{\lambda}) W(j; (X_k)_{k=1}^{\lambda})$ takes a value of the form $w_a w_b$ with probability one. Hence, we find

$$\begin{aligned}
&\mathbb{E}_{X_k, k\neq i, k\neq j}[W(i; (X_k)_{k=1}^{\lambda}) W(j; (X_k)_{k=1}^{\lambda}) \mid f(X_i), f(X_j)] \\
&= \sum_{m=1}^{\lambda-1} \sum_{l=m+1}^{\lambda} w_m w_l \Pr\Bigg[\sum_{k=1}^{\lambda} \mathbb{I}\{f(X_k) \leqslant f(X_i)\} = m \wedge \\
&\qquad\qquad \sum_{k=1}^{\lambda} \mathbb{I}\{f(X_k) \leqslant f(X_j)\} = l \Bigg] ,
\end{aligned}$$

which reads (32). $\square$

Proof of Lemma 3.8. We first prove that $c_f(f(X_i))$ is uniformly distributed in $[0, 1]$. When X_i follows a N-variate normal distribution with a positive definite covariance matrix, one can immediately see that $\Pr[f(X_i) = \alpha] = 0$ for any $\alpha \in \mathbb{R}$. It implies that $f(X_i)$ has a continuous distribution function c_f. Then, from the well-known fact [15, Theorem 2.1], we find that $c_f(f(X))$ is uniformly distributed in $[0, 1]$. This proves the first equality in the lemma statement.

Taking the integral of (31) with respect to $p \in [0, 1]$ and using the formula $\int_0^1 p^{a-1}(1-p)^{\lambda-a} dp = (a-1)!(\lambda-a)!/\lambda!$, we find $\int_0^1 \lambda u_2(x) dx = \sum_{i=1}^{\lambda} w_i^2$. This completes the proof. $\square$

Lemma A.1.

$$\mathbb{E}[u_2(c_f(f(X_i)))^2] \leqslant \frac{1}{\lambda} \sum_{k=1}^{\lambda} w_k^4$$

$$\mathbb{E}[u_3(c_f(f(X_i)), c_f(f(X_j)))^2] \leqslant \frac{2}{\lambda(\lambda-1)} \sum_{k=1}^{\lambda} \sum_{l=k+1}^{\lambda} w_k^2 w_l^2 .$$

Proof of Lemma A.1. From the first paragraph of the proof of Lemma 3.8, we find that $c_f(f(X_i))$ and $c_f(f(X_j))$ are independent and uniformly distributed on $[0, 1]$.

By using Jensen's inequality to exchange the square and the conditional expectation, we have

$$\begin{aligned}
&\mathbb{E}[u_2(c_f(f(X_i)))^2] \\
&\quad= \mathbb{E}_i[\mathbb{E}_{k\neq i}[W(i; (X_k)_{k=1}^{\lambda}) \mid f(X_i)]^2] \\
&\quad\leqslant \mathbb{E}_i[\mathbb{E}_{k\neq i}[W(i; (X_k)_{k=1}^{\lambda})^2 \mid f(X_i)]] \\
&\quad= \mathbb{E}[W(i; (X_k)_{k=1}^{\lambda})^2] \\
&\quad= \sum_{k=1}^{\lambda} w_k^4 \frac{(\lambda-1)!}{(k-1)!(\lambda-k)!} \int_0^1 p^{k-1}(1-p)^{\lambda-k} \mathrm{d}p .
\end{aligned}$$

Using the formula $\int_0^1 p^{k-1}(1-p)^{\lambda-k} \mathrm{d}p = (k-1)!(\lambda-k)!/\lambda!$, the right-most side becomes $\frac{1}{\lambda} \sum_{k=1}^{\lambda} w_k^4$.

Analogously, we obtain

$$\begin{aligned}
&\mathbb{E}[u_3(c_f(f(X_i)), c_f(f(X_j)))^2] \\
&= \mathbb{E}_{i,j}[\mathbb{E}_{k\neq i,j}[W(i; (X_k)_{k=1}^{\lambda}) W(j; (X_k)_{k=1}^{\lambda}) \mid f(X_i) f(X_j)]^2] \\
&\leqslant \mathbb{E}_{i,j}[\mathbb{E}_{k\neq i,j}[W(i; (X_k)_{k=1}^{\lambda})^2 W(j; (X_k)_{k=1}^{\lambda})^2 \mid f(X_i) f(X_j)]] \\
&= \mathbb{E}[W(i; (X_k)_{k=1}^{\lambda})^2 W(j; (X_k)_{k=1}^{\lambda})^2] \\
&= \sum_{k=1}^{\lambda-1} \sum_{l=k+1}^{\lambda} \frac{w_k^2 w_l^2 (\lambda-2)!}{(k-1)!(l-k-1)!(\lambda-l)!} \\
&\qquad \times \int_0^1 \int_0^q 2p^{k-1}(q-p)^{l-k-1}(1-q)^{\lambda-l} \mathrm{d}p\mathrm{d}q .
\end{aligned}$$

Using the formula $\int_0^q p^{k-1}(q-p)^{l-k-1} \mathrm{d}p = (k-1)!(l-k-1)!/(l-1)! q^{l-1}$ and $\int_0^1 q^{l-1}(1-q)^{\lambda-l} \mathrm{d}q = (l-1)!(\lambda-l)!/\lambda!$, the right-most side becomes $\frac{2}{\lambda(\lambda-1)} \sum_{k=1}^{\lambda} \sum_{l=k+1}^{\lambda} w_k^2 w_l^2$. $\square$

Lemma A.2. *The derivative of u_1 is given by*

$$\frac{\mathrm{d}u_1}{\mathrm{d}p} = (\lambda-1) \sum_{k=1}^{\lambda-1} (w_{k+1} - w_k) \binom{\lambda-2}{k-1} p^{k-1}(1-p)^{\lambda-k-1} . \quad (49)$$

Hence, u_1 is Lipschitz continuous, i.e., $|u_1(p) - u_1(q)| \leqslant L|p-q|$, with the Lipschitz constant

$$L \leqslant (\lambda-1) \max_{k\in[\![1,\lambda-1]\!]} |w_{k+1} - w_k| .$$

Proof of Lemma A.2. The derivative of u_1 w.r.t. p is $\lambda \sum_{k=0}^{\lambda-1} w_{k+1} \binom{\lambda-1}{k} \frac{\mathrm{d}}{\mathrm{d}p}[p^k(1-p)^{\lambda-k-1}]$, where

$$\begin{aligned}
\frac{\mathrm{d}p^k(1-p)^{\lambda-k-1}}{\mathrm{d}p} &= kp^{k-1}(1-p)^{\lambda-k-1} \\
&\qquad - (\lambda-k-1)p^k(1-p)^{\lambda-k-2} .
\end{aligned}$$

Substituting them, we have

$$\begin{aligned}
&\frac{\mathrm{d}u_1(p)}{\mathrm{d}p} \\
&= -w_1(\lambda-1)(1-p)^{\lambda-2} + w_\lambda(\lambda-1)p^{\lambda-2} \\
&\quad - \sum_{k=2}^{\lambda-1} w_k \binom{\lambda-1}{k-1} (\lambda-k) p^{k-1}(1-p)^{\lambda-k-1} \\
&\quad + \sum_{k=2}^{\lambda-1} w_k \binom{\lambda-1}{k-1} (k-1) p^{k-2}(1-p)^{\lambda-k} \\
&= (\lambda-1) \sum_{k=2}^{\lambda} (w_k - w_{k-1}) \binom{\lambda-2}{k-2} p^{k-2}(1-p)^{\lambda-k} \\
&= (\lambda-1) \sum_{k=1}^{\lambda-1} (w_{k+1} - w_k) \binom{\lambda-2}{k-1} p^{k-1}(1-p)^{\lambda-k-1} .
\end{aligned}$$

The Lipschitz constant L is the maximum of the absolute value of the above derivative. Note that $\binom{\lambda-2}{k-1} p^{k-1}(1-p)^{\lambda-k-1}$ sums up to one. The absolute value of the right-most side is upper bounded by $(\lambda-1)\max_{k\in[\![1,\lambda-1]\!]} |w_{k+1} - w_k|$. Therefore, $L \leqslant (\lambda-1)\max_{k\in[\![1,\lambda-1]\!]} |w_{k+1} - w_k|$. $\square$

Proof of Lemma 3.9. Since $\mathbf{A}$ is nonnegative definite and symmetric, there exist an orthogonal matrix $\mathbf{E}$ and a nonnegative diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{E}\mathbf{D}\mathbf{E}^{\mathrm{T}}$, where each diagonal element of $\mathbf{D}$ is an eigenvalue of $\mathbf{A}$. Since the distribution of Z is isotropic, Z and $\mathcal{N} = \mathbf{E}^{\mathrm{T}} Z$

follow the same distribution, i.e., $\mathcal{N} = (\mathcal{N}_1, \dots, \mathcal{N}_N)$ is N-multivariate standard normally distributed. Using this fact, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Z^{\mathrm{T}}\mathbf{A}Z}{\mathrm{Tr}(\mathbf{A})} - 1\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N} \frac{[\mathbf{D}]_i}{\mathrm{Tr}(\mathbf{A})}(\mathcal{N}_i^2 - 1)\right)^2\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{[\mathbf{D}]_i[\mathbf{D}]_j}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}\left[(\mathcal{N}_i^2 - 1)(\mathcal{N}_j^2 - 1)\right] \\
&= \sum_{i=1}^{N} \frac{[\mathbf{D}]_i^2}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}\left[(\mathcal{N}_i^2 - 1)^2\right] \\
&\qquad + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} \frac{[\mathbf{D}]_i[\mathbf{D}]_j}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}\left[(\mathcal{N}_i^2 - 1)\right]\mathbb{E}\left[(\mathcal{N}_j^2 - 1)\right] \\
&= \sum_{i=1}^{N} \frac{[\mathbf{D}]_i^2}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}\left[(\mathcal{N}_i^2 - 1)^2\right] = 2\frac{\sum_{i=1}^{N}[\mathbf{D}]_i^2}{\mathrm{Tr}(\mathbf{A})^2} = 2\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} \enspace .
\end{aligned}$$

Here we used the relation $\mathrm{Tr}(\mathbf{A}^2) = \mathrm{Tr}((\mathbf{EDE})^2) = \mathrm{Tr}(\mathbf{D}^2)$. Applying the Schwarz inequality and using Lemma A.1, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|u_2(c_f(f(X)))((Z^{\mathrm{T}}\mathbf{A}Z)/\mathrm{Tr}(\mathbf{A}) - 1)\right|\right]^2 \\
&\quad \leqslant \mathbb{E}[|u_2(c_f(f(X)))|^2]\mathbb{E}[((Z^{\mathrm{T}}\mathbf{A}Z)/\mathrm{Tr}(\mathbf{A}) - 1)^2] \\
&\quad \leqslant 2\left(\frac{1}{\lambda}\sum_{k=1}^{\lambda} w_k^4\right)\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} \enspace . \quad \square
\end{aligned}$$

Proof of Lemma 3.10. Let $\mathbf{A} = \mathbf{EDE}^{\mathrm{T}}$ be the eigen decomposition of $\mathbf{A}$ as we do in the proof of Lemma 3.9, and let $\mathcal{N} = \mathbf{E}^{\mathrm{T}}Z_1$ and $\tilde{\mathcal{N}} = \mathbf{E}^{\mathrm{T}}Z_2$. Then,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Z_1^{\mathrm{T}}\mathbf{A}Z_2}{\mathrm{Tr}(\mathbf{A})}\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N} \frac{[\mathbf{D}]_i}{\mathrm{Tr}(\mathbf{A})}[\mathcal{N}]_i[\tilde{\mathcal{N}}]_i\right)^2\right] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{[\mathbf{D}]_i[\mathbf{D}]_j}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}\left[[\mathcal{N}]_i[\tilde{\mathcal{N}}]_i[\mathcal{N}]_j[\tilde{\mathcal{N}}]_j\right] \\
&= \sum_{i=1}^{N} \frac{[\mathbf{D}]_i^2}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}[[\mathcal{N}]_i^2]\mathbb{E}[[\tilde{\mathcal{N}}]_i^2] \\
&\qquad + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} \frac{[\mathbf{D}]_i[\mathbf{D}]_j}{\mathrm{Tr}(\mathbf{A})^2}\mathbb{E}[[\mathcal{N}]_i]\mathbb{E}[[\tilde{\mathcal{N}}]_i]\mathbb{E}[[\mathcal{N}]_j]\mathbb{E}[[\tilde{\mathcal{N}}]_j] \\
&= \frac{\sum_{i=1}^{N}[\mathbf{D}]_i^2}{\mathrm{Tr}(\mathbf{A})^2} = \frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} \enspace .
\end{aligned}$$

Applying the Schwarz inequality and using Lemma A.1,

$$\begin{aligned}
&\mathbb{E}\left[\left|u_3(c_f(f(X_1)), c_f(f(X_2)))(Z_1^{\mathrm{T}}\mathbf{A}Z_2)/\mathrm{Tr}(\mathbf{A})\right|\right]^2 \\
&\quad \leqslant \mathbb{E}\left[|u_3(c_f(f(X_1)), c_f(f(X_2)))|^2\right]\mathbb{E}\left[((Z_1^{\mathrm{T}}\mathbf{A}Z_2)/\mathrm{Tr}(\mathbf{A}))^2\right] \\
&\quad \leqslant \left(\frac{2}{\lambda(\lambda - 1)}\sum_{k=1}^{\lambda}\sum_{l=k+1}^{\lambda} w_k^2 w_l^2\right)\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} \enspace . \quad \square
\end{aligned}$$

Lemma A.3. *$c_f(f(X)) = c_N(H_N)$.*

Proof of Lemma A.3. The probability of the i.i.d. copy $f(\tilde{X})$ of $f(X)$ being smaller than $f(X)$ is written as

$$\begin{aligned}
&c_f(f(X)) \\
&= \Pr[f(\tilde{X}) < f(X)] \\
&= \Pr[f(\tilde{X}) - f(\boldsymbol{m}^{(t)}) < f(X) - f(\boldsymbol{m}^{(t)})] \\
&= \Pr\left[\frac{f(\tilde{X}) - f(\boldsymbol{m}^{(t)})}{\|\nabla f(\boldsymbol{m}^{(t)})\|^2/\mathrm{Tr}(\mathbf{A})} < \frac{f(X) - f(\boldsymbol{m}^{(t)})}{\|\nabla f(\boldsymbol{m}^{(t)})\|^2/\mathrm{Tr}(\mathbf{A})}\right] \\
&= \Pr\left[\frac{\bar{\sigma}}{c_m}\mathbf{e}^{\mathrm{T}}\tilde{Z} + \frac{\bar{\sigma}^2}{2c_m^2}\frac{\tilde{Z}^{\mathrm{T}}\mathbf{A}\tilde{Z}}{\mathrm{Tr}(\mathbf{A})} < \frac{\bar{\sigma}}{c_m}\mathbf{e}^{\mathrm{T}}Z + \frac{\bar{\sigma}^2}{2c_m^2}\frac{Z^{\mathrm{T}}\mathbf{A}Z}{\mathrm{Tr}(\mathbf{A})}\right] \\
&= \Pr\left[\frac{\bar{\sigma}}{c_m}(\mathbf{Qe})^{\mathrm{T}}\tilde{\mathcal{N}} + \frac{\bar{\sigma}^2}{c_m^2}\frac{\tilde{\mathcal{N}}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\tilde{\mathcal{N}}}{2\,\mathrm{Tr}(\mathbf{A})}\right. \\
&\qquad \left. < \frac{\bar{\sigma}}{c_m}(\mathbf{Qe})^{\mathrm{T}}\tilde{\mathcal{N}} + \frac{\bar{\sigma}^2}{c_m^2}\frac{\tilde{\mathcal{N}}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\tilde{\mathcal{N}}}{2\,\mathrm{Tr}(\mathbf{A})}\right] \\
&= \Pr\left[\frac{\bar{\sigma}}{c_m}[\tilde{\mathcal{N}}]_1 + \frac{\bar{\sigma}^2}{c_m^2}\frac{\tilde{\mathcal{N}}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\tilde{\mathcal{N}}}{2\,\mathrm{Tr}(\mathbf{A})}\right. \\
&\qquad \left. < \frac{\bar{\sigma}}{c_m}[\mathcal{N}]_1 + \frac{\bar{\sigma}^2}{c_m^2}\frac{\mathcal{N}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\mathcal{N}}{2\,\mathrm{Tr}(\mathbf{A})}\right] \\
&= \Pr\left[\tilde{H}_N < H_N\right] = c_N(H_N) \enspace .
\end{aligned}$$

Here we used (28). $\square$

Lemma A.4. *For any $\epsilon > 0$,*

$$\Pr\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|H_N - \left[\frac{\bar{\sigma}[\mathcal{N}]_1}{c_m} + \frac{\bar{\sigma}^2}{2c_m^2}\right]\right| \geqslant \epsilon\right] \leqslant \frac{\bar{\sigma}^4}{2\epsilon^2 c_m^4}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} \enspace .$$

Proof of Lemma A.4. By Chebyshev's inequality, for any $\epsilon > 0$

$$\begin{aligned}
&\Pr\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|H_N - \left[\frac{\bar{\sigma}[\mathcal{N}]_1}{c_m} + \frac{\bar{\sigma}^2}{2c_m^2}\right]\right| \geqslant \epsilon\right] \\
&\quad \leqslant \frac{\bar{\sigma}^4}{4\epsilon^2 c_m^4}\mathbb{E}\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left(\frac{\mathcal{N}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\mathcal{N}}{\mathrm{Tr}(\mathbf{A})} - 1\right)^2\right] \enspace . \quad (50)
\end{aligned}$$

Analogously to what we have done in the proof of Lemma 3.9, we find

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left(\frac{\mathcal{N}^{\mathrm{T}}\mathbf{QAQ}^{\mathrm{T}}\mathcal{N}}{\mathrm{Tr}(\mathbf{A})} - 1\right)^2\right] \\
&\quad = 2\frac{\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\mathrm{Tr}((\mathbf{QAQ}^{\mathrm{T}})^2)}{\mathrm{Tr}(\mathbf{A})^2} = 2\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} \enspace .
\end{aligned}$$

Inserting the above equality to (50), we end the proof. $\square$

Lemma A.5. *For any $\alpha \geqslant 1$,*

$$\mathbb{E}\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}|c_N(H_N) - \Phi([\mathcal{N}]_1)|^{\alpha}\right] \leqslant 3\left[\frac{\bar{\sigma}^2}{4\pi c_m^2}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right]^{\frac{1}{3}} \enspace .$$

Proof of Lemma A.5. The trivial inequality $|c_N(H_N) - c([\mathcal{N}]_1)| \leqslant 1$ for $\alpha \geqslant 1$ gives that $|c_N(H_N) - c([\mathcal{N}]_1)|^{\alpha} \leqslant |c_N(H_N) - c([\mathcal{N}]_1)|$. Then, we have $\mathbb{E}[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}|c_N(H_N) - c([\mathcal{N}]_1)|^{\alpha}] \leqslant \mathbb{E}[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}|c_N(H_N) - c([\mathcal{N}]_1)|]$ and it suffices to show the desired inequality for $\alpha = 1$.

Consider the case $\alpha = 1$. For any $\epsilon > 0$ and any $\boldsymbol{m} \in \mathbb{R}^N \setminus \{x^*\}$,

$$
\begin{aligned}
&|c_N(H_N) - \Phi([\mathcal{N}]_1)| \\
&= |\Pr[\tilde{H}_N < H_N] - \Pr[[\tilde{\mathcal{N}}]_1 < [\mathcal{N}]_1]| \\
&= |\Pr[\tilde{H}_N < H_N] \\
&\qquad - \Pr[(\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2 < (\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2]| \\
&= |\mathbb{E}[\mathbb{I}\{\tilde{H}_N < H_N\}] \\
&\qquad - \mathbb{E}[\mathbb{I}\{(\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2 < (\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2\}]| \\
&\leqslant \mathbb{E}[|\mathbb{I}\{\tilde{H}_N < H_N\} \\
&\qquad - \mathbb{I}\{(\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2 < (\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2\}|] \\
&\leqslant \mathbb{E}[\mathbb{I}\{|((\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2) - ((\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| < \epsilon\} \\
&\qquad + \mathbb{I}\{|H_N - ((\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon\} \\
&\qquad + \mathbb{I}\{|\tilde{H}_N - ((\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon\}] \\
&= \Pr[|((\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2) - ((\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| < \epsilon] \\
&\qquad + \mathbb{I}\{|H_N - ((\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon\} \\
&\qquad + \Pr[|\tilde{H}_N - ((\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon] \\
&= \Pr[|[\tilde{\mathcal{N}}]_1 - [\mathcal{N}]_1| < \epsilon/(\tfrac{\bar{\sigma}}{c_m})] \\
&\qquad + \mathbb{I}\{|H_N - ((\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon\} \\
&\qquad + \Pr[|\tilde{H}_N - ((\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon] \\
&\leqslant \Pr[|[\tilde{\mathcal{N}}]_1 - [\mathcal{N}]_1| < \epsilon/(\tfrac{\bar{\sigma}}{c_m})] \\
&\qquad + \mathbb{I}\{\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |H_N - ((\tfrac{\bar{\sigma}}{c_m})[\mathcal{N}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon\} \\
&\qquad + \Pr[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |\tilde{H}_N - ((\tfrac{\bar{\sigma}}{c_m})[\tilde{\mathcal{N}}]_1 + \tfrac{1}{2}(\tfrac{\bar{\sigma}}{c_m})^2)| \geqslant \tfrac{1}{2}\epsilon] \ ,
\end{aligned}
$$

where the probability and the expectation are taken for $\tilde{H}_N$ and $[\tilde{\mathcal{N}}]_1$. Taking the expectation of the right-most side of the above equality for H_N and $[\mathcal{N}]_1$, we find

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |c_N(H_N) - \Phi([\mathcal{N}]_1)|\right] \\
&\quad\leqslant \Pr\left[|[\tilde{\mathcal{N}}]_1 - [\mathcal{N}]_1| < \epsilon/(\tfrac{\bar{\sigma}}{c_m})\right] \\
&\qquad + \mathbb{E}\left[\mathbb{I}\left\{\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|H_N - \left(\frac{\bar{\sigma}}{c_m}[\mathcal{N}]_1 + \frac{1}{2}\frac{\bar{\sigma}^2}{c_m^2}\right)\right| \geqslant \frac{1}{2}\epsilon\right\}\right] \\
&\qquad + \Pr\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|\tilde{H}_N - \left(\frac{\bar{\sigma}}{c_m}[\tilde{\mathcal{N}}]_1 + \frac{1}{2}\frac{\bar{\sigma}^2}{c_m^2}\right)\right| \geqslant \frac{1}{2}\epsilon\right] \\
&\quad= \Pr\left[\left|[\tilde{\mathcal{N}}]_1 - [\mathcal{N}]_1\right| < \epsilon/(\tfrac{\bar{\sigma}}{c_m})\right] \\
&\qquad + 2\Pr\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|H_N - \left(\frac{\bar{\sigma}}{c_m}[\mathcal{N}]_1 + \frac{1}{2}\frac{\bar{\sigma}^2}{c_m^2}\right)\right| \geqslant \frac{1}{2}\epsilon\right] \\
&\quad= \Pr\left[|[\mathcal{N}]_1| < \epsilon/\left(2^{\frac{1}{2}}(\tfrac{\bar{\sigma}}{c_m})\right)\right] \\
&\qquad + 2\Pr\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}}\left|H_N - \left(\frac{\bar{\sigma}}{c_m}[\mathcal{N}]_1 + \frac{1}{2}\frac{\bar{\sigma}^2}{c_m^2}\right)\right| \geqslant \frac{1}{2}\epsilon\right] \ .
\end{aligned}
$$

The first term on the right-most side is upper bounded as

$$
\begin{aligned}
\Pr[|[\mathcal{N}]_1| < \epsilon/(2^{\frac{1}{2}}(\tfrac{\bar{\sigma}}{c_m}))] &= \int_0^{\epsilon/(2^{\frac{1}{2}}(\frac{\bar{\sigma}}{c_m}))} (2/\pi)^{\frac{1}{2}} \exp(-x^2/2)\mathrm{d}x \\
&\leqslant (2/\pi)^{\frac{1}{2}}(\epsilon/(2^{\frac{1}{2}}(\tfrac{\bar{\sigma}}{c_m}))) = \epsilon/(\pi^{\frac{1}{2}}(\tfrac{\bar{\sigma}}{c_m})) \ .
\end{aligned}
$$

The second term is upper bounded by $\frac{\bar{\sigma}^4}{\epsilon^2 c_m^4}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}$ in light of Lemma A.4. Since the above inequality holds for any $\epsilon > 0$, letting $\epsilon = (2\pi^{\frac{1}{2}}(\frac{\bar{\sigma}}{c_m})^5(\mathrm{Tr}(\mathbf{A}^2)/\mathrm{Tr}(\mathbf{A})^2))^{\frac{1}{3}}$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |c_N(H_N) - \Phi([\mathcal{N}]_1)|\right] \\
&\qquad\leqslant \frac{\epsilon c_m}{\pi^{\frac{1}{2}}\bar{\sigma}} + \frac{\bar{\sigma}^4}{\epsilon^2 c_m^4}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2} = 3\left(\frac{\bar{\sigma}^2}{4\pi c_m^2}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right)^{\frac{1}{3}} \ . \quad \square
\end{aligned}
$$

Proof of Lemma 3.11. In light of Lemma A.2, we find that the function u_1 is Lipschitz continuous, i.e., $|u_1(x) - u_1(y)| \leqslant L|x - y|$ for any $x, y \in [0, 1]$, with the Lipschitz constant L upper bounded by $(\lambda - 1)\max_{k\in[\![1,\lambda-1]\!]}|w_{k+1} - w_k|$. Then, applying the Schwarz inequality and Lemma A.5, we have

$$
\begin{aligned}
&\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |\mathbb{E}[u_1(c_N(H_N))[\mathcal{N}]_1] - \mathbb{E}[u_1(\Phi([\mathcal{N}]_1))[\mathcal{N}]_1]| \\
&\leqslant \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \mathbb{E}[|u_1(c_N(H_N)) - u_1(\Phi([\mathcal{N}]_1))||[\mathcal{N}]_1|] \\
&\leqslant \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \mathbb{E}[|u_1(c_N(H_N)) - u_1(\Phi([\mathcal{N}]_1))|^2]\mathbb{E}[|[\mathcal{N}]_1|^2] \\
&= \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} \mathbb{E}[|u_1(c_N(H_N)) - u_1(\Phi([\mathcal{N}]_1))|^2] \\
&\leqslant \sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} L\mathbb{E}[|c_N(H_N) - \Phi([\mathcal{N}]_1)|^2] \\
&\leqslant L\mathbb{E}\left[\sup_{\boldsymbol{m}\in\mathbb{R}^N\setminus\{x^*\}} |c_N(H_N) - \Phi([\mathcal{N}]_1)|^2\right] \\
&\leqslant 3L\left(\frac{(\bar{\sigma}/c_m)^2}{4\pi}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right)^{\frac{1}{3}} \\
&\leqslant 3(\lambda - 1)\max_{k\in[\![1,\lambda-1]\!]}|w_{k+1} - w_k|\left(\frac{(\bar{\sigma}/c_m)^2}{4\pi}\frac{\mathrm{Tr}(\mathbf{A}^2)}{\mathrm{Tr}(\mathbf{A})^2}\right)^{\frac{1}{3}} \ .
\end{aligned}
$$

Note that $c_f(f(X)) = c_N(H_N)$ (Lemma A.3). This completes the proof. $\square$

Proof of Lemma 3.12. Letting $p_{i:\lambda}$ be the pdf of $\mathcal{N}_{i:\lambda}$, we have

$$
\begin{aligned}
\lambda \int u_1(\Phi([\mathcal{N}]_1))p([\mathcal{N}]_1)[\mathcal{N}]_1 d[\mathcal{N}]_1 &= \sum_{k=1}^{\lambda} w_k \int p_{i:\lambda}([\mathcal{N}]_1)[\mathcal{N}]_1 d[\mathcal{N}]_1 \\
&= \sum_{k=1}^{\lambda} w_k \mathbb{E}[\mathcal{N}_{i:\lambda}] \ . \quad \square
\end{aligned}
$$